Machine Learning
Data Analytics

Friedrich-Alexander-Universität
Technische Fakultät

Balance Assessment and Fall Risk Prediction from Earables using Machine Learning

Master's Thesis in Medical Engineering

submitted
by

Alexander Klingebiel

born 02.01.2000 in Hannover

Written at

Machine Learning and Data Analytics Lab
Department Artificial Intelligence in Biomedical Engineering
Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU)

Advisors:	Ann-Kristin Seifer, M. Sc.
	Sophie Fleischmann, M. Sc.
	Prof. Dr. Bjoern Eskofier
Started:	28.11.2024
Finished:	28.05.2025

Ich versichere, dass ich die Arbeit ohne fremde Hilfe und ohne Benutzung anderer als der angegebenen Quellen angefertigt habe und dass die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegen hat und von dieser als Teil einer Prüfungsleistung angenommen wurde. Alle Ausführungen, die wörtlich oder sinngemäß übernommen wurden, sind als solche gekennzeichnet.
Die Richtlinien des Lehrstuhls für Bachelor- und Masterarbeiten habe ich gelesen und anerkannt, insbesondere die Regelung des Nutzungsrechts.

Erlangen, den 28. Mai 2025

Übersicht

Tragbare Sensoren - Wearables - wie Hörgeräte oder In-Ear-Kopfhörer („Earables“) ermöglichen eine unauffällige Gesundheitsüberwachung älterer Menschen, da sie Bewegungsdaten erfassen können, ohne den Alltag zu beeinträchtigen. Dadurch eignen sie sich für die kontinuierliche Erfassung von Gleichgewicht und Mobilität, welche zentrale Faktoren bei der Einschätzung des Sturzrisikos sind. Für den Einsatz in Klinik oder Alltag müssen Earables zunächst anhand etablierter Standards validiert werden. Diese Arbeit untersuchte, ob aus einer einzelnen ohrgetragenen Inertialmesseinheit (IMU) extrahierte Bewegungsmerkmale zur Vorhersage des Sturzrisikos bei einer gemischten Erwachsenenstichprobe mit 59 Teilnehmenden (∅ 60,1 ±14,3 Jahre) genutzt werden können. Während statischer Gleichgewichtsaufgaben wurden parallel IMU- und Kraftmessplattendaten aufgezeichnet. Eine standardisierte Merkmalsextraktion wurde verwendet, um Zeit-, Frequenz- und Flächenmerkmale aus dem IMU-Beschleunigungssensor und den Drucksignalen (COP) der Kraftmessplatte zu berechnen. Die Kraftmessplatte erfüllte zwei Funktionen: Erstens diente sie als Goldstandard zur Validierung der IMU-basierten Merkmale. Zweitens wurde sie als Referenz für die Sturzrisikovorhersage genutzt, indem maschinelle Lernverfahren (Random Forests, XGBoost, Support Vector Machine) separat auf COP- bzw. IMU-Merkmalen trainiert wurden. Das Sturzrisiko wurde als binäre Zielvariable über die selbstberichtete Falls Efficacy Scale (FES) definiert. Zusätzlich wurde der kognitive Status über den Montreal Cognitive Assessment (MoCA) erfasst. Die technische Validierung zeigte starke Korrelationen im Zeitbereich zwischen IMU- und Kraftplattenmerkmalen, während im Frequenzbereich nur moderate Übereinstimmungen beobachtet wurden. In der Klassifikation erzielten beide Modalitäten nur eingeschränkte Vorhersageleistungen (ROC-AUC bis 0,71; F1 bis 0,47) und zeigten hohe Varianz über Cross-Validation-Folds. Ähnliche Herausforderungen zeigten sich bei der Vorhersage des kognitiven Status anhand der MoCA-Werte. Mögliche Ursachen waren die begrenzte Aussagekraft statischer Aufgaben, die Subjektivität der Zielgrößen und ein unausgewogenes Datenset mit wenigen Hochrisikopersonen. Diese Ergebnisse deuten darauf hin, dass weder IMU- noch kraftmessplattenbasierte Merkmale im vorliegenden Datensatz ausreichten, um das Sturzrisiko zuverlässig vorherzusagen. Die technische Validierung lege nahe, dass Earables wesentliche Bewegungsaspekte des Gleichgewichts im ruhigen Stand erfassen können. Zukünftige Studien sollten longitudinale Datenerhebungen mit dokumentierten Stürzen, dynamischere Aufgaben und standardisierte Protokolle in den Fokus rücken. Die Kombination von IMU-Daten mit weiteren Gesundheitskennwerten könnte zudem die individuelle Risikoprofilierung verbessern. Auch wenn die Klassifikationsergebnisse begrenzt ausfielen, weisen die Ergebnisse der technischen Validierung auf das Potenzial von Earables für eine praxisnahe Gleichgewichtsanalyse im Alter hin.

Abstract

Ear-worn wearables, also known as earables, are emerging as promising tools for unobtrusive health monitoring in older adults. Devices such as hearing aids or earbuds can collect motion data during daily life, enabling continuous assessment of balance and mobility. Since balance impairments and fall risk often increase with age-related or neurological conditions, these technologies offer the potential for early detection and intervention outside clinical environments. However, their ability to capture meaningful movement-related information must first be validated against established standards. This thesis aimed to evaluate whether motion features extracted from a single ear-worn inertial measurement unit (IMU) can be used to predict fall risk in older adult population of 59 participants (∅ 60.1 $\pm$14.3 years). Participants performed static balance tasks while data were simultaneously recorded from an ear-worn IMU and force plates. A standardized feature extraction pipeline was applied to derive time-, frequency-, and area-domain features from the IMU's accelerometer and the force plate's center of pressure (COP) signals. The force plates served two purposes: Firstly, it served as a gold standard to validate the IMU-derived features. Secondly, it provided a reference for evaluating fall risk prediction, as machine learning models (Random Forests, XGBoost, Support Vector Machine) were trained separately on COP-derived features from the force plate and accelerometer-derived features from the IMU. Fall risk was defined as a binary label using the self-reported Falls Efficacy Scale (FES). In addition, cognitive status prediction was explored using Montreal Cognitive Assessment (MoCA) scores. The technical validation revealed strong correlations between IMU and force plate features in the time domain, while frequency-domain features showed moderate correspondence. In the classification task, both modalities showed limited performance in predicting fall risk with ROC-AUC values of up to 0.71 and F1 values of up to 0.47, as well as high variability across cross-validation folds. Similar challenges were found in the prediction of cognitive status using MoCA scores. These limitations may be explained by the limited discriminatory power of static tasks, the subjective nature of the outcome labels, and an imbalanced dataset with relatively few high-risk participants. These findings suggest that neither IMU- nor force plate-derived motion features were sufficient for reliable fall risk prediction using the available dataset and outcome measures. However, the technical validation confirms that earable IMUs can capture aspects of balance-related motion during quiet standing. Future research should prioritize longitudinal data collection with documented fall events, use more dynamic tasks, and adopt standardized protocols. Integrating IMU data with additional health metrics could further support individualized risk profiling. Nevertheless, the strong agreement between IMU and force plate sway features, particularly in the time domain, demonstrates the potential of ear-worn IMUs for practical balance assessment in older adults.

Contents

Chapter 1

Introduction

1.1 Motivation

Maintaining balance is a fundamental aspect of daily living, and impairments in balance are a major risk factor for falls, particularly in older adults. Falls represent a significant public health issue, affecting approximately one-third of adults aged 65 and older each year, and often leading to functional decline, severe injuries, and increased healthcare costs [Wan24] [Rub06]. In addition to physical injuries, recurrent falls in older adults can cause psychological trauma, often leading to a fear of falling. This fear may result in a progressive decline in mobility, as individuals begin to avoid everyday activities, which in turn contributes to further physical deterioration and social withdrawal [Vai20]. Therefore, early and accurate balance assessment is crucial for the prevention of falls and the maintenance of quality of life in aging populations. Traditional clinical assessments of balance, such as the Berg Balance Scale (BBS) [Ber92] and the Timed Up and Go (TUG) Test [Pod91], are widely used due to their simplicity, practicality, and ability to predict fall risk [Man10]. However, these assessments are inherently subjective, often exhibit ceiling effects, and may lack the sensitivity required to detect subtle changes in postural control or early deterioration in balance performance [Man10]. Objective measurements, such as those possible with Force Plates (FPs), offer high precision but are expensive, require specialized equipment and are impractical for widespread use in the community or at home. In addition, their use is limited by the need for trained professionals and lack of portability, making them less suitable for routine clinical examinations or large-scale screening [Noa23b]. To address these limitations, wearable Inertial Measurement Units (IMUs) have emerged as a promising solution for mobile, cost-effective, and objective balance assessment. IMUs are capable of capturing dynamic postural control in a variety of settings, enabling real-world monitoring outside laboratory environments

[Ghi19][Man10]. However, most existing IMU-based systems for balance assessment were attached to the lower back, trunk, or limbs, which can reduce compliance due to discomfort or the visibility of the device, often leading to aversion or embarrassment in public settings [Ghi19] [Vij21]. In contrast, ear-worn devices, known as earables, offer an inconspicuous and socially acceptable alternative and are capable of capturing various physical and physiological phenomena. Positioned at the ear, they provide direct access to head movements and other signals relevant for monitoring body dynamics. Their unobtrusive form factor and integration into widely used consumer devices, such as hearing aids and wireless earbuds, support comfortable and continuous use in everyday life [Röd22]. Despite these advantages, the clinical and technical validity of earable-derived signals for assessing postural control remains underexplored. It is essential to evaluate whether the data captured from earable devices reflect meaningful measures of balance and whether they can contribute to fall risk prediction in practical settings. Moreover, although earable data are increasingly used in research, their clinical utility is still limited by the need for valid and reliable methods, as well as the expertise required to interpret the complex data they generate [Día19] [Azi23]. Applying computational methods, including machine learning, may reveal complex associations in the data that traditional analyses cannot detect [Mus24]. These techniques may enhance the ability to extract relevant information from noisy, real-world signals and support automated, scalable balance assessments, ultimately enabling personalized evaluations and timely interventions. Validating such approaches using earable devices with clinically grounded metrics could enable more targeted monitoring and individualized fall prevention strategies, which is particularly important given the aging population and the increasing burden on healthcare systems.

1.2 Related Work

1.2.1 IMU-based Balance Assessment

The use of IMUs for balance assessment has gained increasing attention as a mobile and objective alternative to traditional laboratory-based methods. Unlike FPs, which offer high precision but are confined to clinical or biomechanical laboratory environments, IMUs enable the measurement of postural sway in everyday settings with minimal setup. Their portability, low cost, and ease of use make them particularly suitable for use in clinical settings, home monitoring, and research involving older adults or individuals with mobility impairments [Man10; Ghi19].

IMUs have been applied in a variety of balance assessment contexts, including gait balance monitoring [Lia23], postural sway analysis [Pol20], and static balance assessment [Noa23a]. A key factor influencing the effectiveness of IMU-based balance assessment is the placement of

the sensor on the body. Since the location determines which aspects of movement are captured, researchers have devoted considerable attention to identifying optimal placements for postural sway analysis. For assessments targeting whole-body stability, sensors are most commonly positioned near the body's center of mass, typically at the waist, lumbar spine, or sternum, where sway is biomechanically most pronounced [God08]. A review by Ghislieri et al. [Ghi19] found that the lower trunk, particularly around the L3–L5 vertebrae, was the most frequently used placement in balance studies due to its effectiveness in capturing trunk acceleration and sway amplitude. Other placements, including the limbs and head, were less common. In addition to sensor location, the review highlighted the importance of task selection, noting that both static and dynamic tasks, such as quiet standing, gait, and sit-to-stand transitions, are employed to probe different dimensions of balance performance. However, the review highlights significant heterogeneity in sensor placement, feature extraction, and validation protocols, with limited attention to head- or ear-mounted configurations.

One of the most influential contributions in this area is the ISway system developed by Mancini et al. [Man12], which introduced a suite of sway metrics derived from a single trunk-mounted IMU during quiet standing tasks. The ISway system demonstrated that time-domain and frequency-domain features such as root mean square acceleration, jerk, and centroidal frequency could reliably differentiate between healthy individuals and those with balance impairments, including Parkinson's disease patients. Crucially, these features showed strong correlations with traditional FP metrics, providing early evidence that IMUs could approximate the biomechanical gold standard in balance assessment.

One application that highlights the practical value of IMU-derived sway metrics is their integration into standardized clinical assessment tools. One prominent example is the Objective Balance Error Scoring System (oBESS) introduced by Brown [Bro13], which aimed to provide a more consistent and objective alternative to the traditional Balance Error Scoring System (BESS). The BESS is a clinical tool for assessing postural stability, in which a professional evaluator scores balance errors during six standardized stance conditions performed with eyes closed and hands on hips on firm and foam surfaces [Fin09]. Although the traditional BESS is simple and inexpensive, it depends heavily on the experience of the evaluators and is known to have low inter-rater reliability. To overcome these limitations, the oBESS system employed IMUs placed on the head, trunk, and limbs to automatically detect movement patterns associated with balance errors. In a sample of healthy adults, the system quantified acceleration and angular velocity across body segments to generate an objective error score, which showed strong agreement with expert visual ratings and reduced the variability associated with subjective scoring. According to Brown,

the oBESS provides a reliable method for quantifying balance and can accurately predict BESS scores in healthy individuals with low total balance errors using only acceleration data from a single IMU positioned at the forehead [Bro13].

Trunk-mounted IMUs have been widely validated for balance assessment, demonstrating strong correlations with FP metrics and clinical scores such as the BBS and TUG when placed near the center of mass. This thesis will investigate whether similar validation can be achieved using ear-worn sensors, which offer a more unobtrusive placement for real-world applications.

1.2.2 Balance Assessment using Earables

Earables have gained attention as unobtrusive tools for capturing physiological and behavioral signals in daily life. They integrate inertial sensors into familiar form factors such as hearing aids, earbuds, or smartglasses, enabling continuous monitoring in naturalistic settings. Ear-worn sensors have been applied in diverse motion analysis contexts, including gait analysis [Jar15], human activity recognition [Bob24], speech and jaw movement detection [Kha21], and sleep posture monitoring [Ngu16]. Röddiger et al. [Röd22] provide a systematic review and taxonomy of sensing phenomena using earables, highlighting their applications in areas such as activity recognition, physiological monitoring, and environmental sensing. Among these, the assessment of postural control and balance is an emerging focus, enabled by the integration of IMUs into consumer devices like hearing aids and smartglasses. Although trunk-mounted IMUs benefit from proximity to the center of mass, earables offer a more discrete and comfortable alternative that may better support long-term and everyday use. Despite listing a wide range of earable applications, the review contains minimal research specifically related to balance or fall risk assessment.

One early validation study in this domain was conducted by Salisbury et al. [Sal18], who evaluated the use of smartglasses equipped with an embedded IMU to measure postural sway. In their work, head-mounted sway metrics obtained from the glasses were compared to those from a conventional waist-mounted IMU during a series of standardized balance stances. The results showed a strong correlation between the two sensor locations, particularly when analyzing total sway power and task-based asymmetries. These findings suggest that head-mounted sensors can serve as viable alternatives to trunk-mounted IMUs for balance assessment, with additional advantages in portability and integration into devices already used in daily life. While the results showed strong correlation with trunk-mounted IMUs, the study did not assess predictive validity for clinically relevant outcomes such as fall risk.

In the pilot study by Saldana et al. [Sal17], researchers explored the feasibility of using a low-cost Virtual Reality Head-Mounted Display (VRHMD) to assess balance in older adults.

The study involved 13 participants aged 65 and older, divided into groups of those at risk of falls and controls. Participants underwent a series of balance assessments using the VRHMD, which included modules designed to evaluate baseline balance, reaction time, and postural control. The VRHMD's submillimeter tracking capabilities allowed for precise measurement of head movements, and its data were compared to traditional FP measurements to assess validity. The findings indicated that the VRHMD could reliably detect differences in balance performance between the two groups, particularly in the rate of anteroposterior tilt changes. However, the study's limitations include a small sample size and the need for further research to generalize the findings to broader populations.

Further work by Gafton et al. [Gra19] demonstrated that a single head-mounted IMU can detect subtle changes in postural sway related to individual characteristics and physiological states. Participants completed static balance trials under various conditions. By analyzing sway power, visual dependency ratios, and weight-bearing asymmetry during static balance tasks, the study showed that head-mounted measurements were sensitive to factors such as sex, substance use, and concussion status. These sway metrics reflect different aspects of postural control: sway power quantifies the overall intensity of body sway, visual dependency ratios compare balance performance with eyes open versus closed to assess reliance on vision, and weight-bearing asymmetry indicates uneven load distribution between the left and right side of the body. Notably, individuals with recent concussions exhibited significantly elevated sway power, indicating impaired neuromotor control. These findings highlight the potential of head-mounted IMUs as practical and responsive tools for identifying balance impairments in real-world or clinical screening contexts.

The studies have shown that head- and ear-worn IMUs can reliably capture sway metrics during static balance tasks, with performance comparable to trunk-mounted sensors in controlled settings. This thesis will explore whether these ear-worn measurements can also serve as a valid input for fall risk classification, expanding their use beyond pure sway quantification and examine how well fall risk predictions based on features from ear-worn IMUs align with those derived from FP data.

1.2.3 Fall Risk Assessment with IMUs

IMUs are widely studied for their potential to assess fall risk in older adults through objective measurement of balance and gait. One of the foundational works in this area is by Howcroft et al. [How13], who systematically reviewed 40 studies that employed inertial sensors to assess fall risk in older adults. Most studies used IMUs positioned near the lower back to capture trunk motion during walking or quiet standing tasks, often under both single- and dual-task conditions. Fallers and non-

fallers were typically classified based on retrospective self-reported fall history or clinical criteria, though definitions varied across studies. The authors found that inertial sensor data could effectively distinguish between fallers and non-fallers, with many studies validating sensor-derived measures against clinical assessments such as the BBS, TUG, or fall history questionnaires. Notably, half of the reviewed studies used predictive models, most commonly regression techniques, but also including classifiers such as decision trees, neural networks, and support vector machines. The review emphasized the need for greater methodological standardization across studies, especially regarding sensor placement, outcome definitions, and the use of prospective validation.

In the context of clinical populations, Ullrich et al. [Ull22] developed machine learning models to predict fall risk in individuals with Parkinson's disease using two weeks of continuous, real-world IMU gait data. IMUs were placed on the lower back, and various gait features were extracted. The outcome label was based on a prospective fall diary over a 6-month follow-up period, classifying participants as fallers or non-fallers. Among several tested models, a Random Forest classifier trained on aggregated daily gait features achieved the highest performance, with a balanced accuracy of 74.0%, sensitivity of 60.0%, and specificity of 88.0%. The study highlighted that real-world gait behavior, as opposed to short clinical assessments, may better reflect actual fall risk in Parkinson's disease patients. While real gait data showed predictive value, the study did not examine sway-based features from static tasks or sensor positions apart from the lower back.

Chen et al. [Che22] reviewed 25 studies applying wearable sensors, primarily IMUs placed at the lower back, waist, or sternum, for fall risk assessment in older adults. Fall risk was typically assessed by using wearable sensor features to predict established clinical indicators of fall risk, such as Timed Up and Go performance, retrospective fall history, or functional mobility scores. In most studies, these clinical tests served as outcome measures or reference standards against which the predictive value of sensor-derived features was evaluated. The review found that wearable sensors can approximate these traditional assessments with reasonable accuracy by extracting features from raw acceleration and gyroscope data, particularly gait variability, stride regularity, and postural sway. While many studies employed statistical comparisons between faller and non-faller groups, the review noted a growing use of supervised machine learning methods to improve classification performance and automate assessment. Overall, the findings support the potential of wearable technologies for fall risk assessment, while highlighting the need to align algorithmic outputs more closely with clinically interpretable outcomes.

A broader distinction between fall risk assessment and fall detection was addressed in the review by Bet et al. [Bet19], which focused using wearable sensors in older adults. Fall detection identifies when a fall has already occurred, typically using real-time sensor data such as sudden

changes in orientation or acceleration. In contrast, fall risk assessment aims to estimate the likelihood of a future fall, either by classifying individuals based on past falls or by evaluating movement and balance independently of fall history. The review identified that fall risk assessment studies primarily employed inertial sensors to monitor gait and balance characteristics, often during standardized walking or static tasks. Fall risk was commonly evaluated using clinical benchmarks such as the TUG, BBS, and fall history questionnaires. These assessments were typically followed by statistical comparisons between faller and non-faller groups, or by developing predictive machine learning models to estimate fall risk. The authors noted that while wearable sensors show good promise in identifying individuals at elevated fall risk, the studies varied widely in terms of sensor placement, number of sensors used, and outcome definitions, making direct comparisons difficult. Moreover, only a few studies validated their results against prospective fall outcomes, underscoring the ongoing need for longitudinal validation and standardized methodologies.

Using the Short Falls Efficacy Scale–International (Short FES-I) to quantify perceived fall risk, Del-Río-Valeiras et al. [del16] examined its association with objective balance performance in older adults with age-related instability. Their findings showed that fear of falling, as reflected in Short FES-I scores, was significantly correlated with both the number of real-life falls and outcomes from computerized dynamic posturography. This supports the relevance of perceived fall risk as a proxy for actual balance impairment. However, while such relationships have been demonstrated using laboratory-based tools, no previous studies have investigated whether sway features from ear-worn inertial sensors can be used to predict fall risk categories defined by the Falls Efficacy Scale-International (FES-I) or Short FES-I. The present work aims to explore this underexamined intersection of wearable sensing, especially with ear-worn devices, and fall risk classification based on a perceived fall assessment tool.

1.3 Purpose

Building on the developments in wearable and earable technologies, the present work aims to systematically validate the use of earable IMUs for balance assessment and fall risk prediction using a previously recorded dataset. The objective is to investigate whether accelerometric features captured from an ear-worn IMU device can serve as reliable indicators of postural stability and predictors of fall risk in older adults. As illustrated in Figure 1.1, the research approach consists of two key validation steps: First, a technical validation is conducted by examining the relationship between Accelerometer (ACC) features from the earable IMU and Center of Pressure (COP) features obtained from a FP during standardized standing balance tasks. This is done by computing

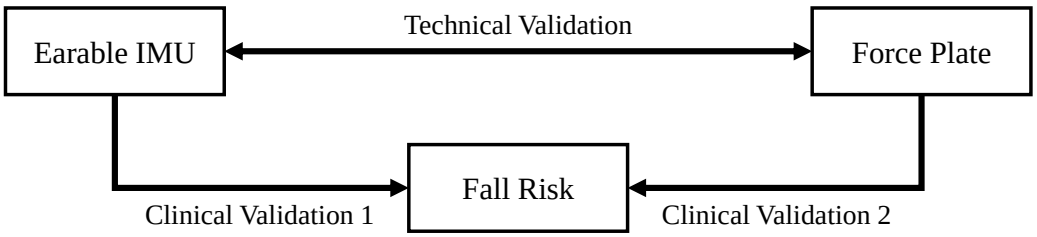


Figure 1.1: Simplified overview of the work packages of this thesis.

Pearson's correlation coefficients and corresponding significance levels across a set of time- and frequency-domain features. This step ensures that the features extracted from the earable accurately reflect biomechanical aspects of postural control traditionally measured using gold-standard laboratory equipment. Second, a clinical validation will be performed by assessing how accurately the FES-I-based classifications of fall risk can be predicted from features extracted from ear-worn IMU and FPs data using machine learning models. This is done by training machine learning classifiers on both IMU and FP features to distinguish between individuals with different levels of perceived fall risk. The performance of these models provides insight into the clinical relevance and predictive value of earable-derived measures. By combining both validation steps, this thesis assesses the accuracy of earable measurements and explores their potential as a practical tool for integrating clinical assessments, such as the FES-I, with wearable technology for real-world use. Demonstrating this potential could support the development of earable-based methods for early identification of individuals at risk of falling and enable broader implementation of fall prevention strategies outside clinical settings.

1.4 Outline

The structure of this thesis is organized as follows: Chapter 2 provides the theoretical background, covering fundamental concepts related to FPs, IMUs, balance assessment tests, and an overview of the selected machine learning approaches and classifiers. Chapter 3 describes the pre-recorded dataset used in this work. Chapter 4 outlines the methods applied, including the correlation analysis between IMUs and FP data as well as the development of a prediction pipeline for fall risk assessment using both sensor modalities. The results and their discussion are presented in Chapter 5. Chapter 6 summarizes the findings, discusses their implications, and highlights the limitations of this work. Finally, Chapter 7 summarizes and concludes the thesis and offers an outlook on future research directions.

Chapter 2

Fundamentals

This chapter presents the fundamental concepts relevant to this thesis. It begins with a brief overview of balance, common impairments, and clinical assessment tools, including both questionnaires and objective tests. The anatomical coordinate system used for biomechanical interpretation is then introduced, followed by an overview of FP measurements and the use of IMUs for balance analysis. Finally, key machine learning concepts are outlined, including common classifiers, data preprocessing, and validation techniques such as nested cross-validation, which are essential for developing reliable predictive models in this thesis.

2.1 Balance and Impairment

Balance is the ability to maintain the body's center of mass over its base of support, relying on the integration of sensory input from the visual, vestibular, and proprioceptive systems [Pol00] [Hor96]. Disruptions in any of these systems can lead to balance impairments, manifesting as dizziness, unsteadiness, or falls [Stu08]. Common causes include inner ear disorders, neurological conditions such as Parkinson's disease and multiple sclerosis, musculoskeletal issues like arthritis, and side effects from certain medications [Agr13] [All13] [Cam18] [Lev12] [Woo09]. The prevalence of balance impairments among older adults is substantial and shows a clear age-related increase. Approximately 30% of individuals aged 65 years and older report experiencing balance impairments or dizziness at some point in their lives [Wan24]. Given the high prevalence and multifactorial nature of balance impairments in older adults, accurate assessment is essential for identifying deficits and guiding interventions. Balance assessment involves a combination of self-report questionnaires and objective clinical tests that evaluate both static and dynamic postural control. Common questionnaires include the Activities-specific Balance Confidence (ABC) Scale,

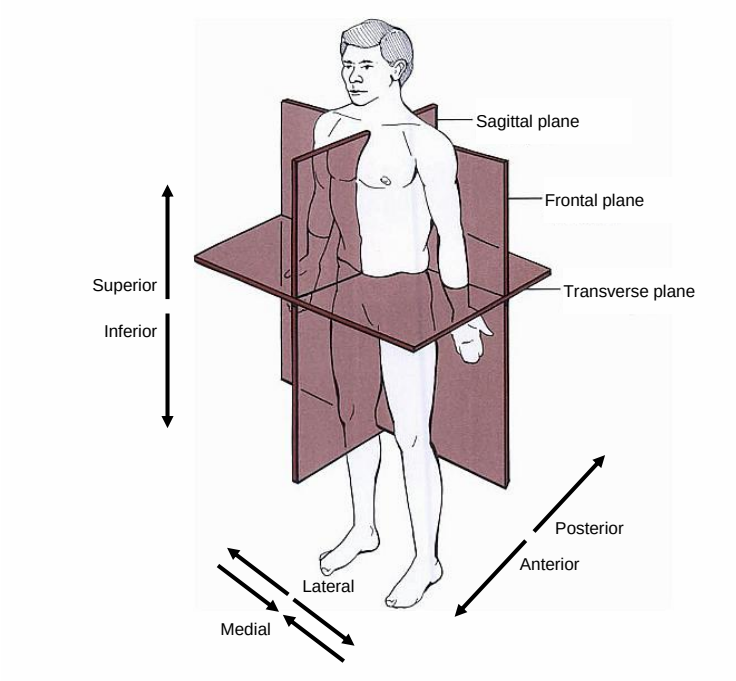


Figure 2.1: Anatomical terms of location labeled on the anatomical position (adapted from [Mad08]).

the Dizziness Handicap Inventory (DHI), and the FES-I, which assess perceived confidence and the impact of dizziness or fear of falling on daily activities [Nno15] [Yar05]. Objective tools like the BBS and the BESS are used to assess static balance, while dynamic balance can be evaluated through the TUG test and the Short Physical Performance Battery (SPPB), which reflect mobility and functional capacity in older adults [Nno15] [Ive13] [Pav16].

2.2 Medical Coordinate System

In the context of anatomical and biomechanical studies, standardized directional terms and body planes are utilized to describe the orientation and movements of body structures precisely. These conventions are fundamental for ensuring consistent communication and are summarized in Table 2.1 and Figure 2.1 [Wer23]. These directional terms are defined relative to the subject's standard anatomical position, which describes the body standing upright, facing forward, with arms at the sides and palms facing outward. This ensures consistency in spatial descriptions regardless of the subject's actual posture [Whi05].

Table 2.1: Standard anatomical terms used to describe positional and directional relationships in the human body.

Medical Term	Translation
Anterior	Front
Posterior	Back
Medial	Towards the midline of the body
Lateral	Away from the midline of the body
Superior	Toward the head / upper
Inferior	Toward the feet / lower

2.3 Balance Assessment Tools

2.3.1 Balance Assessment Instruments

Force Plates

Force Plates (FPs) are precision biomechanical instruments used to quantify Ground Reaction Forces (GRFs) and torques generated during human movement and posture and are commonly used in the assessment of postural stability. They are commonly employed in both static and dynamic posturography to monitor the control of upright stance, where static posturography involves the analysis of quiet erect posture, and dynamic posturography assesses the individual's response to externally applied disturbances [Dua10]. The Center of Pressure (COP), derived from GRFs and moments around the plate's surface, represents the point of application of the resultant force vector on the support surface and reflects the body's neuromuscular responses to maintain equilibrium [Dua10] [Qui21]. Typically, the COP is measured along two axes: the Medio-Lateral (ML) and Antero-Posterior (AP) directions, which correspond to the horizontal components of postural sway [Che21]. The COP is mathematically computed using the measured force and moment components according to the following equations, where F_x, F_y, F_z are the ground reaction forces in the AP, ML, and Superior-Inferior (SI) directions, respectively, M_x, M_y are the moments around the ML and AP axes, and h is the height offset from the force measurement surface to the actual point of contact (e.g., a surface covering) [Dua10]:

$$\text{COP}_{AP} = \frac{-M_y - h \cdot F_x}{F_z}, \quad \text{COP}_{ML} = \frac{M_x - h \cdot F_y}{F_z} \tag{2.1}$$

When the subject is standing barefoot directly on the FP surface, the vertical offset h is effectively zero [Sei12]. Multi-axis FPs measure the three orthogonal force components (F_X, F_Y, F_Z) and the corresponding moments (M_X, M_Y, M_Z), enabling high-resolution tracking of COP trajectories.

These COP trajectories can be visualized as stabilograms, which show the time series of the COP in the AP and ML directions, or as statokinesigrams, which plot the COP in the ML direction against the AP direction in a 2D spatial map. Both visualizations provide insight into postural control strategies and individual stability margins [Dua10]. Quantitative COP metrics, such as total path length, sway area and mean velocity are widely used to assess postural control capacity. For example, increased COP velocity or sway area is generally interpreted as reduced stability and is frequently observed in populations with impaired balance, such as older adults [Qui21].

Inertial Measurement Unit

Inertial Measurement Units (IMUs) are widely used for assessing balance and postural stability in clinical and real-world settings. IMUs typically consist of accelerometers and gyroscopes measuring the inertial acceleration and angular rotation respectively [Ahm13]. Accelerometers are particularly crucial, as they measure linear acceleration across three orthogonal axes, namely AP, ML, and SI, which correspond to the primary directions of sway during standing or walking [Zha22] [Sei23]. The data, usually expressed in units of gravitational acceleration (g), can be converted to standard SI units (m/s^2) by multiplying by 9.80665. This quantitative output enables precise monitoring of sway amplitude, frequency, and variability, which are key parameters in evaluating balance performance. Such sensors are often placed on the lower back, sternum, or even behind the ear to capture whole-body movements with minimal intrusion [Ghi19] [Ata11]. These sensors are essential in various applications including activity recognition, fall detection, gait analysis, and long-term health monitoring [Sli19] [Ghi19]. Wearable accelerometers have been successfully integrated into devices such as hearing aids, wristbands, and waist clips, enabling unobtrusive and continuous tracking [Sei23] [Ata11]. The sampling rate of an IMU depends on the activity of interest, typically between ten and several hundred Hz. Higher sampling rates enable the system to acquire signals with higher precision and frequency, resulting in more precise models at the expense of higher energy consumption [Zha22]. Moreover, their small size, low cost, and ease of integration into wearable systems such as hearing aids or belt clips enhance their utility for continuous, at-home balance [Ata11] [Ghi19]. However, challenges such as placement variability can impact accuracy and minimize errors due to sensor alignment and drift [Ata11] [Ahm13].

2.3.2 Questionnaires and Tests

Falls Efficacy Scale

The Falls Efficacy Scale-International (FES-I) is a standardized instrument used to assess the fear of falling by quantifying an individual's concerns during everyday activities with a 16-item

questionnaire each rated on a scale of one to four [Yar05]. Each item is rated on a four-point scale ranging from 1 (not at all concerned) to 4 (very concerned), resulting in scores that add up to a value between 16 and 64, with higher scores indicating greater fear of falling. The questionnaire covers a broad range of physical and social activities, including walking on uneven surfaces, reaching overhead, bathing, and attending social events. This wide scope makes the FES-I particularly suitable for both frail individuals and more active older adults [Yar05] [Del10]. It has become a standard tool in fall risk assessment, with scores of 16 to 21 indicate low fall risk, while scores of 22 and above indicate higher fall risk. This threshold enables clinicians to differentiate between individuals with lower and higher levels of fear, thus guiding targeted fall prevention strategies [Del10].

Montreal Cognitive Assessment

The Montreal Cognitive Assessment (MoCA) is a brief screening tool widely used to detect mild cognitive impairment, which has implications for balance assessment as cognitive deficits can adversely affect postural control. The MoCA evaluates a wide range of cognitive domains using tasks such as short-term memory recall, visuospatial skills, executive function, attention and working memory, language, and orientation. The total score ranges from 0 to 30, with higher scores indicating better cognitive function. Studies have shown that individuals scoring below 26 on the MoCA often exhibit impairments that may translate into balance and mobility challenges. Thus, a MoCA score of 26 serves as the threshold, distinguishing between cognitively normal subjects and those at risk of balance-related issues [Nas05].

2.4 Machine Learning

Machine learning refers to the ability of computational models to learn from task-specific training data, enabling them to automatically build analytical models and solve related problems. Traditional machine learning techniques such as decision trees [Qui86], Random Forests (RFs) [Bre01], Extreme Gradient Boosting (XGBoost) [Che16], and Support Vector Machines (SVMs) [Cor95] are commonly used as initial approaches for biomedical data analysis. They are particularly effective with smaller datasets, such as those with fewer than 300 samples, and generally offer greater interpretability compared to deep learning models [Zan24] [Bin24].

Figure 2.2 presents a standard machine learning classification pipeline split into training and testing phases. The process begins with data preprocessing and feature extraction to transform raw inputs into useful representations. The prepared data is then divided into a training set, used

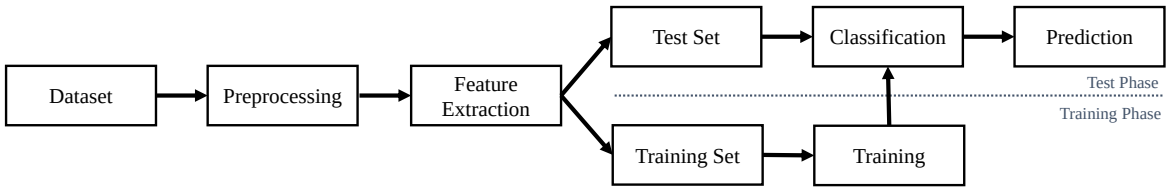


Figure 2.2: The basic modules of a model classification system [Nie83].

for model learning, and a test set for evaluating classification model performance on unseen data [Nie83].

2.4.1 Training and Evaluation

To build a robust model, it is essential to utilize as many available samples as possible during training, while ensuring that the test set is sufficiently large to provide a reliable estimation of predictive performance. If the available dataset has only a small size for training and testing, this can pose a significant challenge. A widely adopted solution is cross-validation, which allows the entire dataset to be systematically used for both training and testing in different iterations [Bis06]. One of the most commonly used variants is k-fold cross-validation, in which the dataset is divided into k equally sized folds. In each of the k iterations, one fold is reserved for validation, while the remaining $k-1$ folds are used for training. This procedure ensures that every data point is used for validation exactly once and for training $k-1$ times. The performance metrics are then averaged over all folds to obtain a more reliable estimate of the model's predictive performance [Ras21].
In most machine learning workflows, it is not enough to just train a model. The process also includes selecting the best hyperparameters and applying preprocessing steps such as scaling or imputation. If these optimization steps and the evaluation of model performance are carried out using the same data, biased performance estimates may result. This problem is particularly prominent in scenarios with small samples, where it is often not possible to create a separate, independent test set. To mitigate this problem, **nested cross-validation** has emerged as a robust method for simultaneously performing model selection and performance estimation with minimal bias. It also enables the systematic comparison of different machine learning algorithms under identical validation conditions. This approach involves two levels of cross-validation: an outer loop, which is reserved for evaluating the generalization performance of the final model, and an inner loop, which is used for training, preprocessing, and hyperparameter tuning. By strictly separating model selection from performance estimation, nested cross-validation provides an

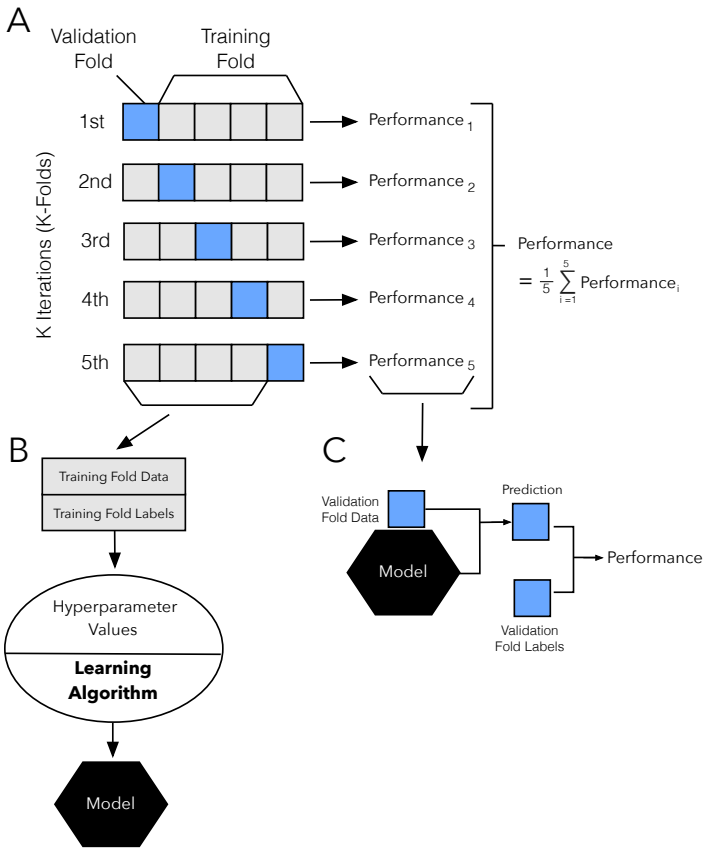


Figure 2.3: The outer cross-validation loop procedure [Ras21].

almost unbiased estimate of prediction error and is particularly well suited for small datasets [Ras21]. The following subsections describe the role of each loop.

Model Evaluation and Validation

In the outer loop of nested cross-validation, the focus lies on evaluating the generalization performance of the final model. To achieve this, k-fold cross-validation is used. The dataset is split into k outer folds, and in each iteration, one fold is held out as an independent test set. The remaining $k-1$ folds are used for training and model optimization to find the best performing model in the inner loop. This best performing model is then tested on the held out independent test set. This strict separation ensures that performance evaluation is conducted on data that has not been used in any part of the training or tuning process. By averaging the results across all outer folds, an unbiased estimate of the model's expected generalization error is obtained. An overview of this process is illustrated using a 5-fold cross-validation scheme in Figure 2.3.

To obtain unbiased performance estimates, especially in imbalanced scenarios, **stratified testing** ensures that the class distribution in the test set mirrors that of the full dataset. When applied in k-fold cross-validation, stratification positively affects the variance and bias of the performance estimates by maintaining consistent class proportions in each fold [Ras21].

Training and Hyperparameter Tuning

Within each outer training loop iteration, an inner cross-validation loop is executed to perform training, preprocessing, and hyperparameter tuning. The inner dataset, which corresponds to the portion of the data not held out by the outer loop, is further split into k folds. In each inner iteration, one fold is used for validation, and the remaining folds are used for training the model with particular hyperparameter configurations. This training process is repeated using a hyperparameter optimization approach, and the configuration yielding the best average performance across inner folds is selected. Examples for these hyperparameter optimization approaches are techniques such as bayesian optimization, randomized search, or grid search, that systematically evaluate various hyperparameter configurations. The final model, trained on the full inner training set using these optimal settings, is then evaluated on the outer loop test fold, as described above. In this way, the inner loop supports model selection, while the outer loop ensures independent validation. This process is illustrated in Figure 2.4. In this particular instance, a 5x2 configuration is employed, meaning that a 5-fold cross-validation is used in the outer loop and a 2-fold cross-validation is applied in the inner loop [Ras21].

When dealing with real-world data, preprocessing of the data has to be applied, including imputation, centering, and scaling strategies. These preprocessing steps are embedded into the cross-validation pipeline and must be applied within each outer loop iteration independently to avoid data leakage and biased estimates. **Imputation** techniques are commonly used to address often occurring missing features. The basic principle of imputation is that the value of a feature that is not present in a particular observation can be approximated using the available data [Saa07]. Methods can be categorized into univariate techniques, which treat each variable independently, often using measures of central tendency, and multivariate approaches considering the relationships among variables to yield more robust estimates. Univariate methods are straightforward and computationally efficient, but they may overlook the underlying structure of the data and introduce bias in subsequent analyses. In contrast, multivariate imputation techniques, such as iterative algorithms and nearest-neighbor methods, model the dependencies between variables and iteratively refine estimates to better approximate the joint distribution. This flexible strategy allows the imputation process to be tailored to the specific characteristics of a dataset, thereby

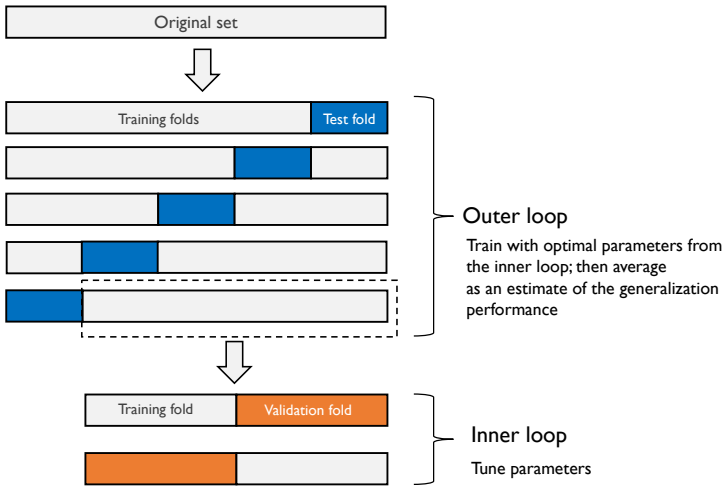


Figure 2.4: The nested cross-validation procedure[Ras21].

enhancing the validity and reliability of statistical inferences drawn from incomplete data [Van12]. Centering and scaling are fundamental preprocessing techniques for input variables. **Centering** involves subtracting the mean from each observation so that the resulting variable has a zero mean. **Scaling** follows by dividing each value by the standard deviation, thereby standardizing the variable to a unit standard deviation. An alternative scaling approach is the min-max method, which rescales the variable to a fixed range, typically [0, 1], by subtracting the minimum value and dividing by the range (maximum minus minimum). These procedures are typically employed to enhance the numerical stability of subsequent computations [Kuh13]. These steps such as imputation and scaling are always fitted exclusively on the training data of a given loop iteration, whether in the inner or outer loop, and then applied to the corresponding validation or test fold to prevent data leakage [Wil23].

2.4.2 Common Machine Learning Classifiers

In the following, several machine learning methods are presented in the context of supervised learning classification, which means that the algorithm is trained on data with correct output labels to infer a mapping function that can then be used to classify unseen examples accurately [Cha21b].

Decision Trees

Machine learning classification algorithms divide the feature space into different decision regions, each corresponding to a specific output class. Decision trees partition the feature space by recursively splitting it into smaller regions [Bis06]. The process begins at the root node, which

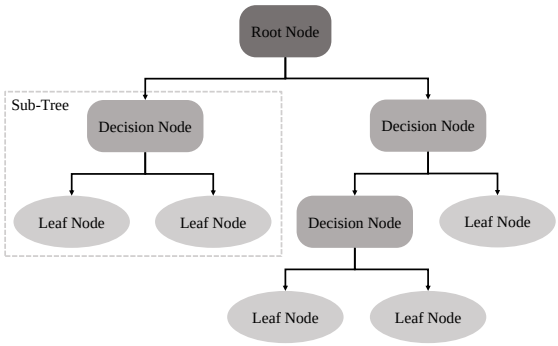
represents the entire feature space. At each decision node, the algorithm selects a feature and a corresponding threshold that best separates the data based on a criterion such as information gain, Gini impurity, or entropy [Cha21b]. The first split divides the data into two groups according to whether the selected feature's value is above or below the threshold. Each of these groups form a sub-tree, which is then further split at new decision nodes, with the algorithm recursively selecting features and thresholds based on the remaining uncertainty. This hierarchical, recursive process continues until a stopping criterion is met. The result is a tree structure, as depicted in Figure 2.5a, in which each leaf node corresponds to a distinct region of the feature space, and each region is assigned a prediction based on the majority class of the training data that falls into that region. This method allows decision trees to model complex decision boundaries in a straightforward and interpretable way [Bis06]. Decision trees are able to handle both categorical and numerical data and highlights important features through a tree-like structure, yet they can be unstable with small data changes and biased towards features with many levels [Bin24].

Random Forest

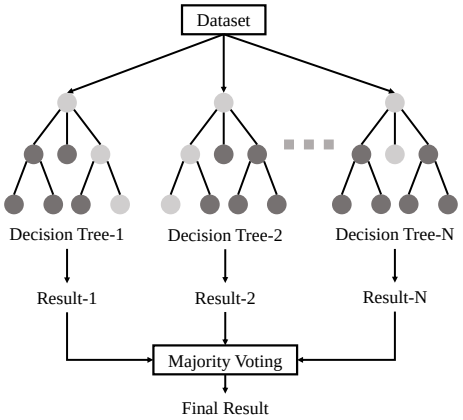
Random Forests (RFs) are an ensemble method that extends decision trees by building multiple trees and aggregating their predictions. Each tree is grown using a bootstrapped sample of the data, which introduces variability among the trees. At every split in a tree, only a random subset of features is considered, reducing correlation between trees and enhancing diversity [Kuh13]. This combination of bootstrapping and random feature selection helps to mitigate overfitting compared to a single decision tree [Bre01]. The ensemble then aggregates the predictions, typically through majority voting for classification to produce a final output. An example of this process is illustrated in Figure 2.5b. This aggregation effectively reduces variance without a significant increase in bias, resulting in improved generalization performance. RFs are robust to outliers and noisy data, but they cannot reliably extrapolate beyond the range of the training data and can be computationally expensive [Bin24].

Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a decision tree ensemble that constructs an additive model by iteratively minimizing a regularized objective function using gradient-based optimization. In this process, first- and optionally second-order derivatives of the loss function are used to guide each step, reducing prediction error while controlling model complexity to prevent overfitting. In contrast to random forests, which build trees independently and then aggregate their predictions, XGBoost constructs trees sequentially, each one aiming to correct the errors of its predecessors. It



(a) An illustration of a decision tree structure [Sar21].



(b) An illustration of a random forest composed of multiple individual decision trees used for classification [Sar21].

also employs techniques like shrinkage (also known as learning rate), maximum depth constraints, and random subsampling of training instances and features to further enhance its predictive performance and reduce overfitting [Ben21]. An important characteristic of XGBoost is its ability to handle sparse input data through sparsity-aware split finding, which enables the model to learn optimal default split directions for missing or zero-valued features, thereby avoiding the need for data imputation or exclusion of incomplete samples. Overall, XGBoost delivers high predictive accuracy and scalability, though its complexity in hyperparameter tuning and reduced interpretability compared to simpler models can be challenging [Che16] [Fri01].

Support Vector Machine

Support Vector Machines (SVMs) are a robust supervised learning technique used for classification, where the goal is to find an optimal hyperplane that maximally separates data points of different classes by enlarging the margin between them. The margin refers to the distance between the hyperplane and the closest data points from each class, known as support vectors, which are critical in defining the decision boundary. Maximizing this margin is essential because a larger margin generally leads to a lower generalization error, making the SVM more robust to noise in the training data [Sar21]. A key aspect of SVMs is the kernel function, which allows the algorithm to implicitly map input features into a higher dimensional space using kernel functions, such as linear, polynomial, radial basis function, or sigmoid, without explicitly computing the transformation, enabling the modeling of non-linear relationships efficiently [Nob06]. This process is visualized in Figure 2.6. SVMs effectively handle high-dimensional data and perform well in both linear and

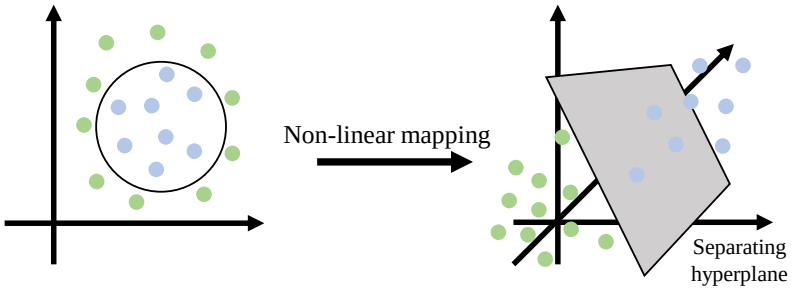


Figure 2.6: Illustration of feature transformation and separating hyperplane [Cha21a].

non-linear classification tasks. However, they require careful scaling and normalization, and may overfit when trained on complex data or small datasets. [Bin24].

Chapter 3

Dataset: Ear-Balance

The data used in this work was prerecorded as part of the clinical study EmpkinS C03 [Emp25] conducted with the University Hospital Erlangen, funded by the German Research Foundation (DFG). The study was originally focused on balance impairments in individuals with Parkinson's disease. In this thesis, the dataset is utilized for the purpose of fall risk prediction in older adults. This chapter describes the characteristics of the test participants, the study design and the data collected from FP and IMU measurements during static balance tasks.

3.1 Population

A total of 70 participants took part in the study, which collected data from individuals diagnosed with Parkinson's disease and age-matched controls. Participants with Parkinson's disease were included if they had a clinical diagnosis of a Parkinson's syndrome according to the guidelines of the German Society of Neurology (Hoehn & Yahr stage I–III), were over 18 years of age, able to stand safely without assistance, and able to walk 4 × 10 meters independently. Additional requirements included the ability to speak and read, as well as providing informed consent. Exclusion criteria comprised aphasia or alexia, visual impairments preventing reading, regular use of walking aids, decompensated cardiopulmonary conditions, a maximum walking distance of less than 100 meters, inability to understand the study procedures, and severe musculoskeletal disorders significantly affecting mobility. Age-matched controls were eligible if they were over 18 years old and had no diagnosed musculoskeletal or neurological disorders. For this work's purpose of balance assessment eleven participants needed to be excluded due to technical issues or invalid tests. This resulted in 59 participants, whose characteristics are summarized in Table 3.1. They were included if they were able to perform at least one static task with valid sensor data from the FP and IMU

	All
Sex (m/f)	39/20
Age (years)	60.1±14.3
Height (cm)	173.6±9.7
Weight (kg)	80.5±17.3
AC/PD	31/28
FES-I	19.8±6.1
FES (low/high)	48/11
Timed-up-and-Go (s)	10.2±3.3
SPPB (points)	9.9±2.1
Fried Frailty Index	0.5±0.8
MoCA	25.8±3.5
MoCA (normal/impaired)	37/22

Table 3.1: Demographic and clinical characteristics of the included participants (N=59). PD = Parkinson's disease, AC = age-matched control. Falls Efficacy Scale-International (FES-I): low fear of falling < 23, high fear ≥ 23 [Yar05]. MoCA: scores ≥ 26 considered cognitively normal, < 26 as impaired [Nas05].

Task	**N**
EC closed	56
EC pad	47
EC semi	48
EC tandem	22
EC wide	56
EO closed	57
EO pad	56
EO semi	54
EO tandem	41
EO wide	55
Overall	**59**

Table 3.2: Number of participants with valid data per balance task with Eyes Closed (EC) and Eyes Open (EO). Overall indicates the total number of participants (N=59), who completed at least one task with valid data.

as well as valid FES-I and MoCA tests. If any of these tests were invalid, they were excluded to ensure adequate comparability between the FP and IMU. The number of participants per task can be seen in Table 3.2.

3.2 Study Design and Procedure

The participants gave written informed consent prior to the recording and the study was approved by the local ethics committee (Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany) Re-No. 20-473-B. Aside from the participant's characteristics, e.g., age, sex and weight, data was collected with clinical questionnaires and instruments. The participants answered the questionnaires of FES-I, MoCA and Fried Frailty Index, which is a standardized questionnaire to assess frailty on a scale of 0 to 5. The TUG and SPPB were performed by the participants, which are clinical tests to evaluate the participants' mobility. The participants were asked to perform five specific balance tasks with varying difficulties while standing, as shown in Figure 3.1. These consisted of standing with feet together (closed), standing with feet shoulder-width apart (wide), standing in a straight line with one foot directly in front of the other (tandem), positioning the heel

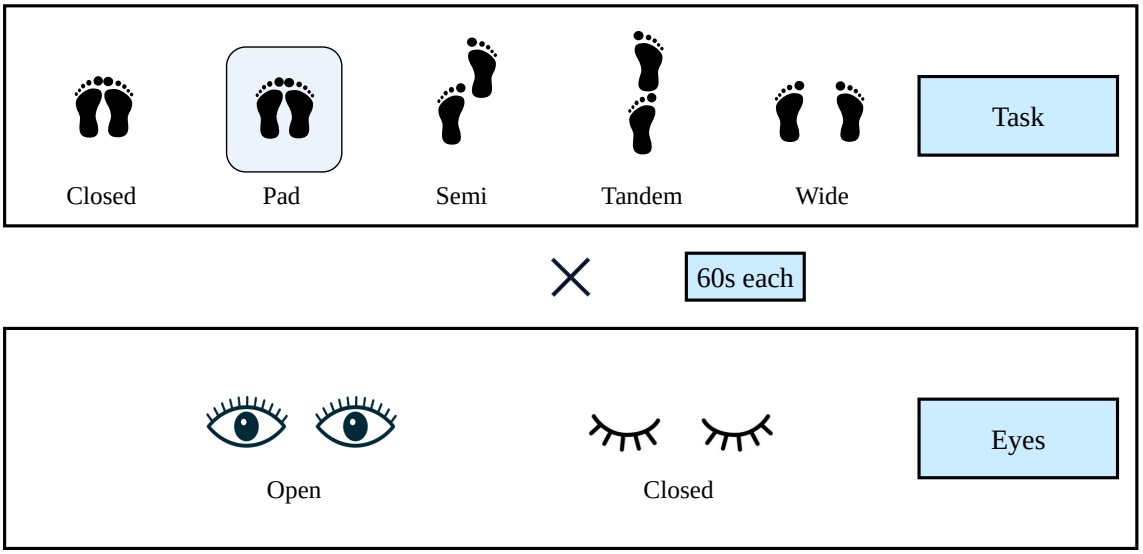


Figure 3.1: Overview of the five examined static balance tasks: closed, pad (foam surface), semi-tandem, tandem, and wide stance. Each task was performed under two visual conditions, with eyes open and eyes closed, resulting in a total of 10 conditions per participant. Each condition lasted 60 seconds.

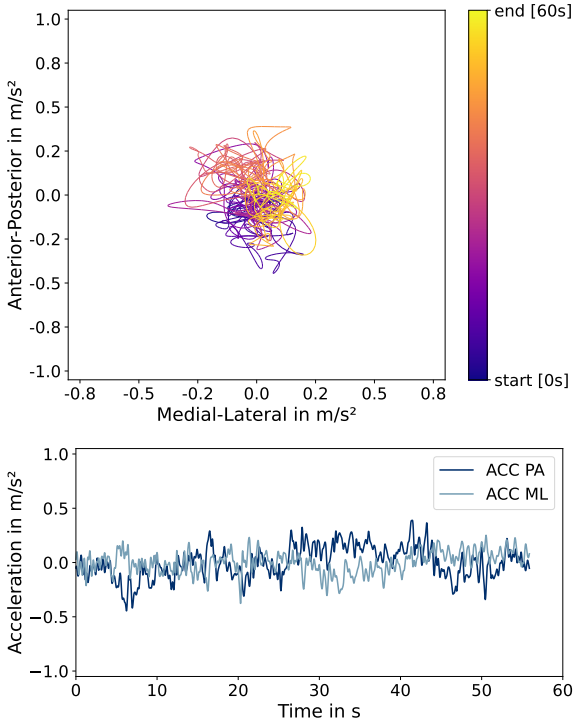
of one foot next to the big toe of the opposite foot (semi), and standing on a foam pad with feet together (pad). Each balance task was performed under two conditions: with EO and with EC. Participants were asked to maintain the position for 60 seconds. If a participant lost balance, the recording was ended early. In addition, they were asked to perform a walking task at their own chosen speed and in 10-meter bouts, which was repeated until they hit the FPs three times with each foot. Data was recorded with FPs on which the participants stood and with a IMU integrated into a hearing aid that the participants were asked to wear, while performing the tasks. The order of the tasks was randomized to avoid sequence effects. The pad and semi tasks were performed on one FP, while the closed, tandem, and wide tasks were executed with the participants standing on two FPs.

3.3 Data of IMUs and Force Plates

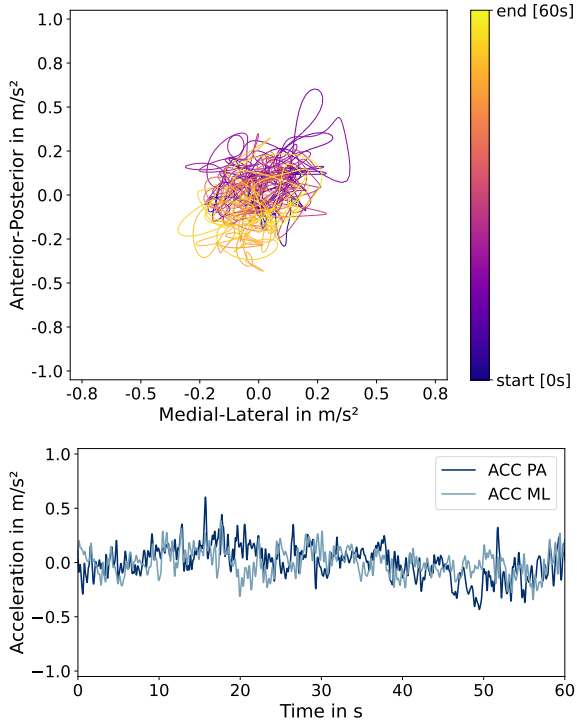
The original study collected data with optical marker tracking, inertial sensors attached to the participant's bodys and two FPs embedded in the floor, to measure GRFs. Additionally, data from an IMU integrated into a hearing aid worn on the right ear was recorded. For this work, the data of the FPs and ACC of the IMU were used. GRFs were measured using two FPs (Bertec, Columbus, USA). Each FP recorded the three orthogonal force components (F_x, F_y, F_z) and the

corresponding moments (M_x, M_y, M_z) at a sampling rate of fs = 1000 Hz. From these signals, the COP was derived to quantify postural sway. The data was initially collected in .c3d files and subsequently converted into .mot files. To ensure consistency across trials with varying participant orientations, the force data was rotated into the participant-specific anatomical coordinate system, aligning the COP trajectories with the anteroposterior and mediolateral axes of the body. IMU data was collected using a Bosch BMI270 sensor, which is built into a hearing aid device (Signia RIC Pure 312) worn on the right ear. The sensor includes a three-axis accelerometer and gyroscope (tri-axial accelerometer ± 2 g; tri-axial gyroscope ± 1000 °/s), and data was sampled at 50 Hz (or 25 Hz in some cases due to technical difficulties) via a smartphone app. The approach for dealing with different sampling rates is described in chapter 4.1. The data of the IMU was saved as .mat files. Due to technical limitations during data acquisition, the recordings from the FPs and the IMU are not perfectly synchronized. This discrepancy arises from the manual labeling of task start and stop times.

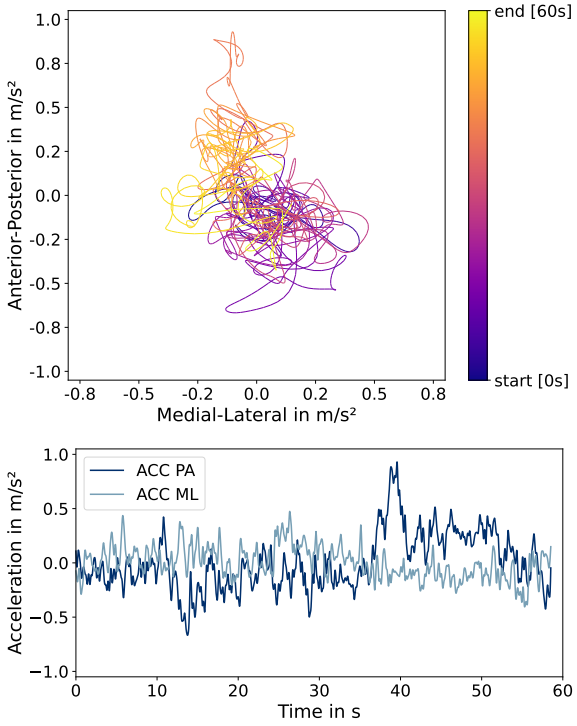
The following two Figures 3.2 and 3.3 illustrate posturography data derived from Force Plate (FP) (Center of Pressure (COP)) and ear-worn Inertial Measurement Unit (IMU) (Accelerometer (ACC)) measurements during a 60-second quiet standing task performed on a foam pad. Figure 3.3 shows the data from the FPs, while Figure 3.2 presents the corresponding IMU data. Each figure is organized to compare two participants: one with low fall risk (Participant 05, FES-I = 16) and one with high fall risk (Participant 33, FES-I = 26), under two sensory conditions: EO and EC. In both figures, the top plots of each subfigure show the statokinesigrams color-coded over time and the bottom plots show the corresponding stabilograms. A visual comparison across conditions and participants reveals that sway amplitude and irregularity are greater in the eyes-closed condition and for the participant with higher fall risk. This pattern is evident in both the COP and ACC data, indicating that the more challenging sensory condition (EC) and the higher fall risk might be associated with reduced postural stability, in in accordance with the analysis of Howcraft et al. [How17]. All data were preprocessed using the pipeline described in Section 4.1.



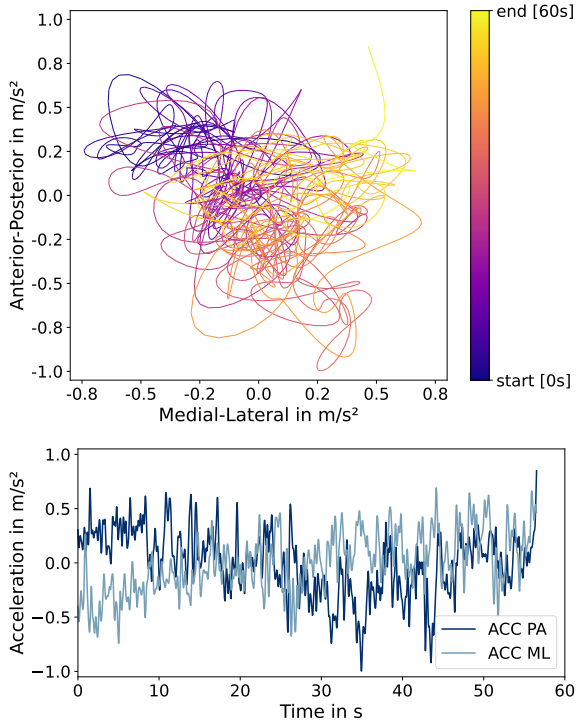
(a) EO, Low Fall Risk



(b) EO, High Fall Risk

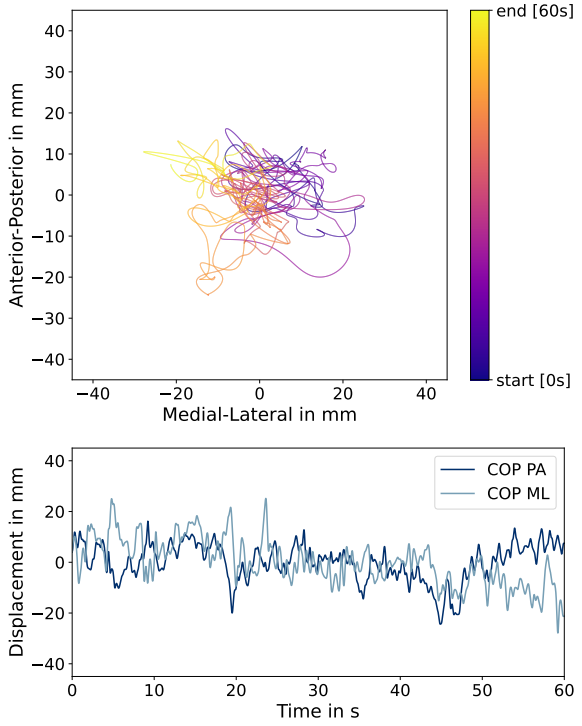


(c) EC, Low Fall Risk

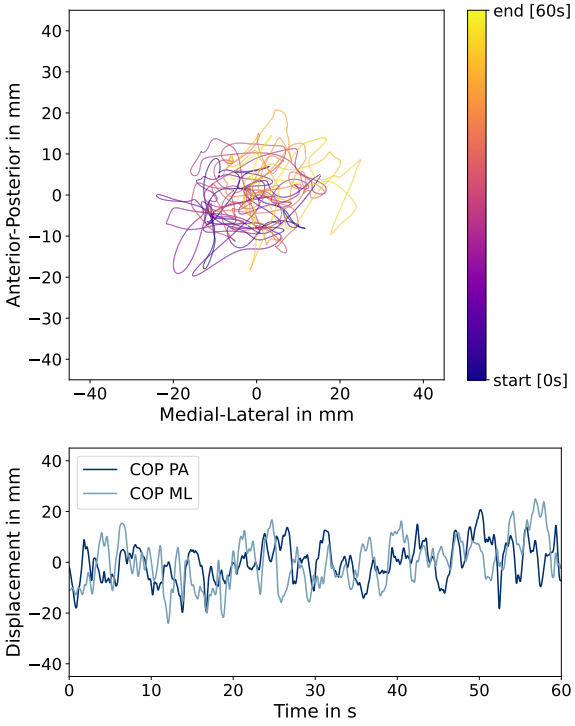


(d) EC, High Fall Risk

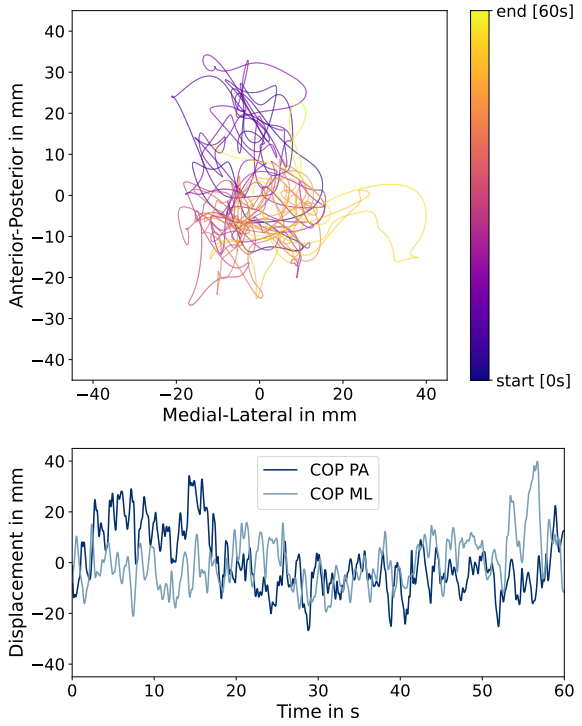
Figure 3.2: Posturography from IMU data for a low fall risk (Pat 5, FES-I = 16) and a high fall risk participant (Pat 33, FES-I = 26) during quiet standing on a foam pad. Subfigures (a) and (b) show measurements under EO conditions; (c) and (d) show the same under EC conditions. Each subfigure displays a statokinesigram (top) and the corresponding stabilogram (bottom) for ML and AP sway of the ACC.



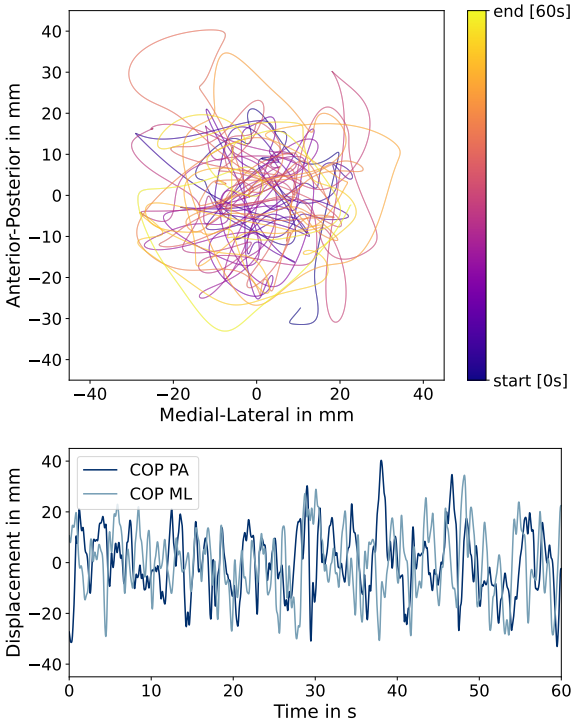
(a) EO, Low Fall Risk



(b) EO, High Fall Risk



(c) EC, Low Fall Risk



(d) EC, High Fall Risk

Figure 3.3: Posturography from FP data for a low fall risk (Pat 5, FES-I = 16) and a high fall risk participant (Pat 33, FES-I = 26) during quiet standing on a foam pad. Subfigures (a) and (b) show measurements under EO conditions; (c) and (d) show the same under EC conditions. Each subfigure displays a statokinesigram (top) and the corresponding stabilogram (bottom) for ML and AP COP displacement.

Chapter 4

Methods

This chapter describes how the data obtained from the earable IMU was used to assess fall risk, which is illustrated in Figure 4.1. First, an overview of the pre-processing procedure is given, followed by a description of the extracted features from the ACC data of the IMU and the COP data of the FP. In the subsequent section, the derived features of ACC and COP are correlated to technically validate the use of earable IMUs. Then, a machine learning pipeline is proposed that predicts the fall risk using the FES-I as ground truth labels first from the ACC features and then from the COP features. A similar approach for predicting cognitive function using MoCA as ground truth labels is then presented. Finally, the evaluation metrics are presented to quantify the correlation and prediction results.

4.1 Pre-Processing

To properly work with IMU and FP data, the pre-recorded dataset was preprocessed. The data of the IMU was originally intended to be recorded at 50 Hz. However, due to technical issues, 8 participants were unintentionally recorded at a lower sampling rate of 25 Hz. Instead of resampling the data to a common frequency, all pre-processing and feature extraction steps were performed using each participant's respective sampling rate. This approach ensured that all derived features were calculated correctly and consistently, regardless of the original sampling rate. To enable a meaningful comparison across participants, the IMU data were aligned to the gravity vector, thereby transforming the sensor signals into the medical coordinate system. This ensures that the vertical axis consistently corresponds to gravity, and that the AP and ML directions are comparable across all recordings. This alignment was performed using the $\mathrm{TrimMeanGravityAlignment}$, function from the *eargait* python library [Sei23], which estimates the gravity vector by computing

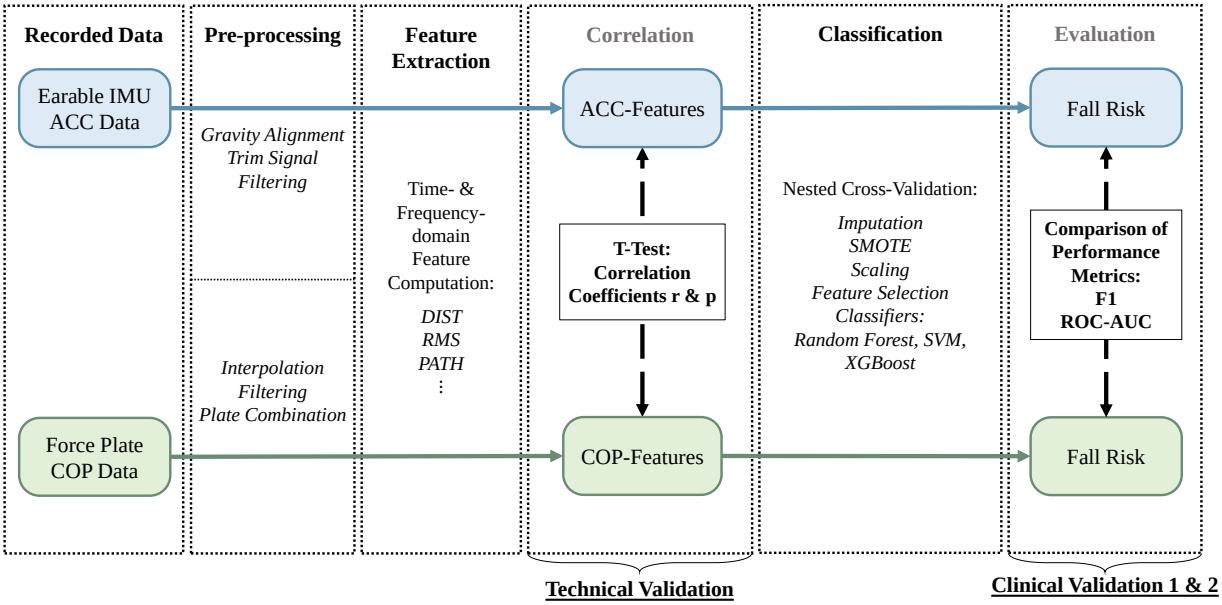


Figure 4.1: Overview of the Data Processing, Correlation, and Classification Workflow using IMU and FP Data.

a trimmed mean of the ACC signal. This method was chosen over the alternative StaticWindowGravityAlignment [Sei23] because the study protocol did not include dedicated motionless periods. As a result, the assumption of a stationary window required for the static method was not met, making the trimmed mean approach a more robust choice under these recording conditions. To minimize the error that could have occurred when the recording of the IMU data were started and stopped manually, two seconds were discarded at the beginning and end of the recording. The data from the head-mounted ACC from each measurement was low-pass filtered using a fourth-order Butterworth filter with a cut-off frequency of 3.5Hz [Man12]. In this work, the acceleration ACC_{ML} in the medio-lateral direction and the acceleration ACC_{AP} in the antero-posterior direction were used to calculate the IMU features.

The FP raw data consists of GRFs (F_x, F_y, F_z), moments/torques (M_x, M_y, M_z), and COP coordinates (COP_x, COP_y). In trials that required two FPs, data were recorded separately with FP A and FP B. Due to errors or measurement inaccuracies, a linear interpolation of some missing data points in the FP data was performed. The data from the FP from each measurement was low-pass filtered using a fourth-order Butterworth filter with a cut-off frequency of 10 Hz [Cub24][Man12]. To determine the global COP locations in the x- and y-directions, the COP positions recorded by both plates, denoted as $x_{\text{COP},a}, x_{\text{COP},b}$ and $y_{\text{COP},a}, y_{\text{COP},b}$, were weighted based on the vertical forces measured by each plate $F_{z,a}, F_{z,b}$. The total vertical force across both plates is given by F_z. The

global COP positions are then computed as follows [Exe11]:

$$COP_{\text{ML}} = \left(x_{\text{COP},a} \times \frac{F_{z,a}}{F_z}\right) + \left(x_{\text{COP},b} \times \frac{F_{z,b}}{F_z}\right) \tag{4.1}$$

$$COP_{\text{AP}} = \left(y_{\text{COP},a} \times \frac{F_{z,a}}{F_z}\right) + \left(y_{\text{COP},b} \times \frac{F_{z,b}}{F_z}\right) \tag{4.2}$$

In this work, the global center of pressure COP_{ML} in the medio-lateral direction and the global center of pressure COP_{AP} in the antero-posterior direction were used to calculate the FP features. Due to the manual labeling of the start and end times of the tasks during data acquisition, the recordings of the FP and the IMU were not perfectly synchronized. As a result, the duration of the COP recordings is, on average, longer than that of the IMU data, with a mean difference of 0.8 seconds ($\pm$3.4 s). If the discrepancy between the recordings for a task was greater than 10 seconds, the subject's task was excluded from the analysis.

4.2 Feature Extraction

The assessment of postural control commonly relies on quantifiable balance features that characterize the variability and frequency content of body sway. Both COP data from the FPs and ACC signals from the IMU capture the dynamic fluctuations of a subject's posture during static tasks. Fundamental time-domain features and frequency-domain features are computed using established signal processing techniques (e.g., lowpass filtering, calculation of Euclidean distances, and fast Fourier transforms). These techniques have proven sensitive and reliable for quantifying postural control in both clinical and experimental settings [Man12] [Pri96]. Despite the inherent differences between COP and ACC signals - namely, that COP data represent displacement while ACC measurements capture acceleration - the computation of most balance features remains analogous across these modalities. Both types of signals are influenced by the same underlying neuromuscular control mechanisms governing postural stability. Provided that appropriate preprocessing (e.g., filtering and coordinate alignment) is applied, the mathematical operations used to extract variability and spectral characteristics from COP data can be directly transferred to ACC data. An important exception applies to the Mean Velocity (MV) feature, which is inherently based on displacement and thus requires a different computation method for ACC signals than for COP data. The validity of this approach is supported by Mancini et al. [Man12] demonstrating that inertial sensors placed on the trunk can yield balance features that closely correlate with those derived from FP measurements [Qui21].

4.2.1 Time-Domain Features

To analyze and use the data of the IMU and the FP in a more meaningful way, features must be extracted to assess human movement and balance. Here, x_i and y_i denote the coordinates at time point i in the Antero-Posterior (AP) and Medio-Lateral (ML) directions, respectively, corresponding to ACC_{ML} and ACC_{AP} for the Accelerometer (ACC) of the Inertial Measurement Unit (IMU), and COP_{ML} and COP_{AP} for the Center of Pressure (COP) of the Force Plates (FPs). The center of the trajectories in each direction $\bar{x}$ and $\bar{y}$ is denoted as, with N being the being the number of samples in the trajectory:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{4.3}$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{4.4}$$

The radius of each point r_i, defined by the Euclidean distance from the center $(\bar{x}, \bar{y})$, is given by:

$$r_i = \sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2} \tag{4.5}$$

Mean Distance from the Center of the Trajectory

The Mean Distance from the Center of the Trajectory (DIST) can be computed by the mean of the euclidean distance of each point to the center of the trajectories [Qui21].

$$\text{DIST} = \frac{1}{N}\sum_{i=1}^{N} |r_i| \tag{4.6}$$

Root Mean Square

The Root Mean Square (RMS) of the radius r_i, which represents the Euclidean distance of each point from the center of the trajectory, is given by [Qui21]:

$$\text{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} r_i^2} \tag{4.7}$$

Sway Path

The Total Sway Path Length (PATH), representing the total trajectory length of the COP or ACC,

is given by sum of the distances between consecutive points [Qui21]:

$$\text{PATH} = \sum_{i=1}^{N-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2} \tag{4.8}$$

Range of Displacement

The Range of Displacement (RANGE) is defined as the maximum Euclidean distance between any two points on the stabilogram [Qui21]:

$$\text{RANGE} = \max_{1 \leq i \leq j \leq N} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{4.9}$$

The computation can be optimized by first determining the convex hull of the point set, as the maximum distance is always found between two of its vertices. By then calculating the pairwise distances among these vertices only, the number of necessary comparisons is significantly reduced compared to evaluating every possible point pair.

Mean Velocity

Mean Velocity (MV) is the only feature computed differently for COP and ACC data. The mean velocity of the COP can be defined as the ratio of the PATH traveled by the COP to the duration of the trial T [Qui21]:

$$\text{MV}_{COP} = \frac{\text{PATH}}{T} \tag{4.10}$$

To compute the MV of the IMU, the ACC signals in the x_i and y_i directions are first integrated over time to yield velocity. Since numerical integration typically introduces low-frequency drift, each integrated signal undergoes a high-pass filtering operation, denoted by $\mathcal{HP}[\cdot]$, to remove this drift. This can be done using a fourth-order high-pass butterworth filter with a cutoff-frequency of 0.01 Hz. Subsequently, the Euclidean norm of the filtered velocity is calculated at each sample, and these values are then averaged over the entire trial duration T using a discrete summation [Man12]:

$$\text{MV}_{ACC} = \frac{1}{T} \sum_{i=0}^{T} \sqrt{\left(\mathcal{HP}\left[\int x_i\right]\right)^2 + \left(\mathcal{HP}\left[\int y_i\right]\right)^2} \tag{4.11}$$

Mean Frequency

To compute the Mean Frequency (MF) of the COP and ACC data, the concept of repeated circular like sway movements around the center of the trajectory is used. It is assumed that postural sway traces an imaginary circle, with the average distance to the center represented by the metric DIST,

serving as the radius of this circle. The Total Sway Path Length (PATH) reflects the accumulated distance traveled along this imaginary trajectory over the entire trial. To estimate how many such loops the body completes, the total path length is divided by the circumference of the circle, and this value is then divided by the total trial duration T, resulting in the number of loops per second [Man12]:

$$\text{MF} = \frac{\text{PATH}}{2 * \pi * \text{DIST} * T} \tag{4.12}$$

Sway Area per Second

The Sway Area per Second (AREA) quantifies the average area traced by the COP or ACC trajectory over the course of a trial. Conceptually, each pair of consecutive points and the mean position form triangular sections whose combined area is then normalized by the trial duration, resulting in an average area per unit time. The centered trajectory can be described as (x_i, y_i) and (x_{i+1}, y_{i+1}) as consecutive points of the trajectory and subdivided into triangles using the shoelace formula. An extra term can be added in practice to connect the last sample back to the first, thus closing the polygon [Qui21]:

$$\text{AREA} = \frac{1}{2T} \sum_{i=1}^{n-1} |x_{i+1} y_i - x_i y_{i+1}| \tag{4.13}$$

4.2.2 Frequency-Domain Features

In addition to time-domain features, frequency-domain characteristics were extracted to capture the distribution of sway dynamics across different spectral components. To quantify the spectral content of the signals, the Power Spectral Density (PSD) was estimated using Welch's method with 10-second segments, 50% overlap, and linear detrending [Qui21] and its coefficients are denoted by $\Gamma_k^X = \Gamma^X(f_k)$, where of X corresponding to the frequency $f_k = k\frac{f_s}{N}$ for $k = 1, \ldots, \frac{N}{2}$ if N is even and $k = 1, \ldots, \frac{N-1}{2}$ if N otherwise. N is the number of samples during the trial and f_s is the sampling rate. The frequency domain measurements for the ACC and COP are limited to the frequency range between 0.15 Hz and 5.0 Hz to capture the dynamics of postural sway. Very low frequency components below 0.15 Hz can excluded as they correspond to slow fluctuations with limited relevance for postural control and may be influenced by non-physiological drift or sensor noise [Pri96]. The frequency range from $f_{\text{inf}} = 0.15\,\text{Hz}$ to $f_{\text{sup}} = 5\,\text{Hz}$ corresponds to the indices:

$$k_{\text{inf}} = \left\lfloor \frac{f_{\text{inf}}\, N}{f_s} \right\rfloor + 1 \quad \text{and} \quad k_{\text{sup}} = \left\lfloor \frac{f_{\text{sup}}\, N}{f_s} \right\rfloor,$$

Moreover, the ℓ-th moment of the PSD can be denoted by:

$$M_\ell^X = \sum_{k=k_{\text{inf}}}^{k_{\text{sup}}} f_k^\ell \, \Gamma_k^X, \tag{4.14}$$

Total Power

The Total Power (PWR) can be obtained by first computing the Euclidean distance to form a single time series, and then transforming this signal to the frequency domain. Summing Γ_k^X over the relevant frequency bins (k_{inf} to k_{sup}) then yields the overall energy content of the trajectory.

$$\text{PWR} = \sum_{k=k_{\text{inf}}}^{k_{\text{sup}}} \Gamma_k^X \tag{4.15}$$

Median Frequency

The Median Frequency (F50) is defined as the frequency at which the cumulative sum of the PSD reaches 50% of the total power, effectively serving as the median frequency of the spectrum [Qui21].

$$\text{F50} = \inf\left\{ k^* \in \mathbb{N} \;\middle|\; \sum_{k=k_{\text{inf}}}^{k^*} \Gamma_k^X \;\geq\; 0.5 \sum_{k=k_{\text{inf}}}^{k_{\text{sup}}} \Gamma_k^X \right\} \times \frac{f_s}{N} \tag{4.16}$$

95% Power Frequency

The 95% Power Frequency (F95) is the frequency below which 95% of the total spectral power is contained and thus defines the upper limit of the energy distribution of the signal [Qui21]:

$$\text{F95} = \inf\left\{ k^* \in \mathbb{N} \;\middle|\; \sum_{k=k_{\text{inf}}}^{k^*} \Gamma_k^X \;\geq\; 0.95 \sum_{k=k_{\text{inf}}}^{k_{\text{sup}}} \Gamma_k^X \right\} \times \frac{f_s}{N} \tag{4.17}$$

Centroidal Frequency

The Centroidal Frequency (CF) quantifies the location of the spectral mass within the PSD and provides a single value that represents the "center of gravity" of the energy distribution. It is defined as the square root of the ratio of the second spectral moment to the zeroth moment [Qui21]:

$$\text{CF} = \sqrt{\frac{M_2^X}{M_0^X}} \tag{4.18}$$

Table 4.1: Overview of extracted features for ACC and COP data.

Short Name	Full Name / Description	ACC Unit	COP Unit
DIST	Mean distance from center of trajectory	m/s^2	mm
RMS	Root mean square distance from center	m/s^2	mm
PATH	Total sway path length	m/s^2	mm
RANGE	Maximum displacement between any two points	m/s^2	mm
MV	Mean velocity	m/s	mm/s
MF	Mean frequency (loops per second relative to trajectory)	Hz	Hz
AREA	Sway area per second	m^2/s^5	mm^2/s
PWR	Total power (sum of spectral power)	m^2/s^4	mm^2
F50	Median frequency (50% power frequency)	Hz	Hz
F95	95% power frequency	Hz	Hz
CF	Centroidal frequency	Hz	Hz
FD	Frequency dispersion (spread of spectral content)	-	-

Frequency Dispersion

Frequency Dispersion (FD) measures the spread or variability of the spectral content of the PSD. Ranging from zero, indicating that the energy is concentrated at a single frequency, to one, which would correspond to a uniform distribution across the examined frequency band, this feature provides valuable information about the heterogeneity in the frequency content of the signal [Qui21].

$$\text{FD} = \sqrt{1 \; - \; \frac{\left(M_1^X\right)^2}{M_2^X \, M_0^X}} \tag{4.19}$$

4.3 Correlation between Earable IMU and Force Plate Features

As FPs are already an established part of clinical assessment, the aim of this method was to validate ear-worn IMUs by correlation with established extracted features, as depicted in Figure 4.2. After pre-processing the ACC signals of the IMU and COP signals of the FP, a series of features were extracted for the time-domain (DIST, RMS, RANGE, MV, MF, AREA) and frequency-domain (PWR, F50, F95, CF, FD) using the equations and implementations described in Section 4.2 and summarized again with the respective units in Table 4.1. For each combination of static balance task and extracted feature, the corresponding values from the IMU and FP were subjected to Pearson correlation analysis. This yielded one r coefficient and associated p value per task–feature pair, quantifying the linear relationship between the two modalities. The r- and p-value are described

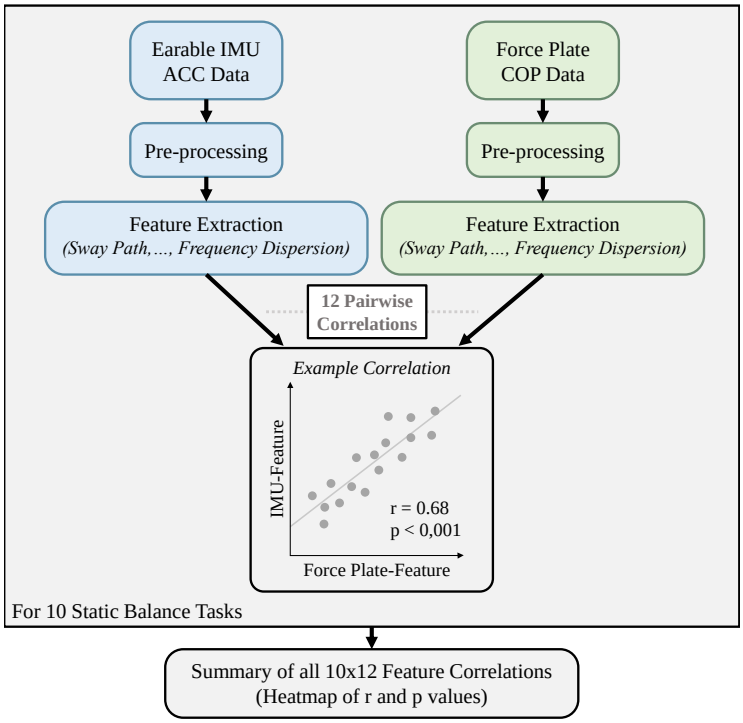


Figure 4.2: Correlation of ACC and COP features using Pearson's r and p-values.

in Section 4.6. These statistics quantify the strength and significance of the relationship between the modalities and thus provide an assessment of the validity of the IMU compared to the gold standard FP measurements. Figure 4.2 illustrates the overall workflow, from signal pre-processing through feature extraction to correlation analysis resulting in a 10x12 heatmap for both correlation coefficients.

4.4 Fall Risk Classification

Earable IMU Features

To classify participants by fall risk, a supervised machine learning framework was implemented using sway features derived from the ACC of a single ear-worn IMU. Detailed information on the characteristics of the pre-recorded dataset can be found in the Chapter 3, with the features derived from time-, area- and, frequency domains for each static balance task, as shown in Chapter 4.2. The features were aggregated per participant to form one representative feature vector. The binary target variable was defined based on the Falls Efficacy Scale-International (FES-I)), where a threshold of 22 was used to distinguish between low fall risk (FES-I ≤ 22, N = 48) and high fall risk (FES-I

> 22, N = 11) [Del10]. To ensure robust and unbiased model evaluation, a nested cross-validation strategy with a 5×4 split was employed, balancing the need for reliable performance estimation with sufficient data in each fold given the limited sample size. The outer loop used 5-fold stratified cross-validation to evaluate model generalization. For hyperparameter optimization, a 4-fold inner loop was combined with 300 iterations of randomized search. The machine learning pipeline was structured to overcome various challenges stemming from the pre-recorded dataset, including missing data (from uncompleted tasks), class imbalance and high feature dimensionality. The processing steps of the features included imputation of missing values, oversampling of the minority class, feature scaling, and dimensionality reduction. The final classification models used in the outer loop included Extreme Gradient Boosting (XGBoost), Random Forest (RF), and Support Vector Machine (SVM)), each with their own set of tuned hyperparameters. Different combinations of classifiers and imputation strategies were evaluated, with performance metrics averaged across the outer folds of the nested cross-validation. An overview of the entire classification pipeline, from features to final prediction of FES-I class, is shown in Figure 4.3.

This conceptual pipeline was implemented using Python, and the specific libraries, processing steps and model configurations are described below. For each of the aggregated feature vectors per participant, an additional binary feature was added for classification to indicate task completion per condition, while retaining information about whether a particular balance task was completed or aborted, allowing imputation strategies without loss of information about the aborted tasks. For handling missing data itself, several imputation strategies from the *scikit-learn*[1] library were evaluated. These included univariate approaches using *SimpleImputer* with *'mean'* and *'median'*, as well as multivariate methods such as *KNNImputer*. Separate classification runs were conducted for each combination of classifier and imputation strategy. The classifiers used were *XGBClassifier* from the *xgboost* library, and *RandomForestClassifier* and *SVC* from the *scikit-learn* library. Within each of these runs, the full machine learning pipeline was constructed using the *ImbPipeline* class from the *imbalanced-learn* library. Several preprocessing components were included as tunable steps in the pipeline. Class imbalance in the target variable was addressed through the Synthetic Minority Over-sampling Technique (SMOTE) implementation from *imbalanced-learn*[2] library. Feature scaling was performed using either the *StandardScaler* or the *MinMaxScaler*, both from *scikit-learn*. Dimensionality reduction was achieved using *SelectKBest* from *scikit-learn*, which applies Analysis of Variance (ANOVA) F-statistics to retain the most informative features. SMOTE, Scaler, and *SelectKBest* were treated as hyperparameters and optimized jointly in the inner cross-validation loop. Additionally, each classifier's model-specific parameters were

[1] https://scikit-learn.org/stable/index.html
[2] https://imbalanced-learn.org/stable/

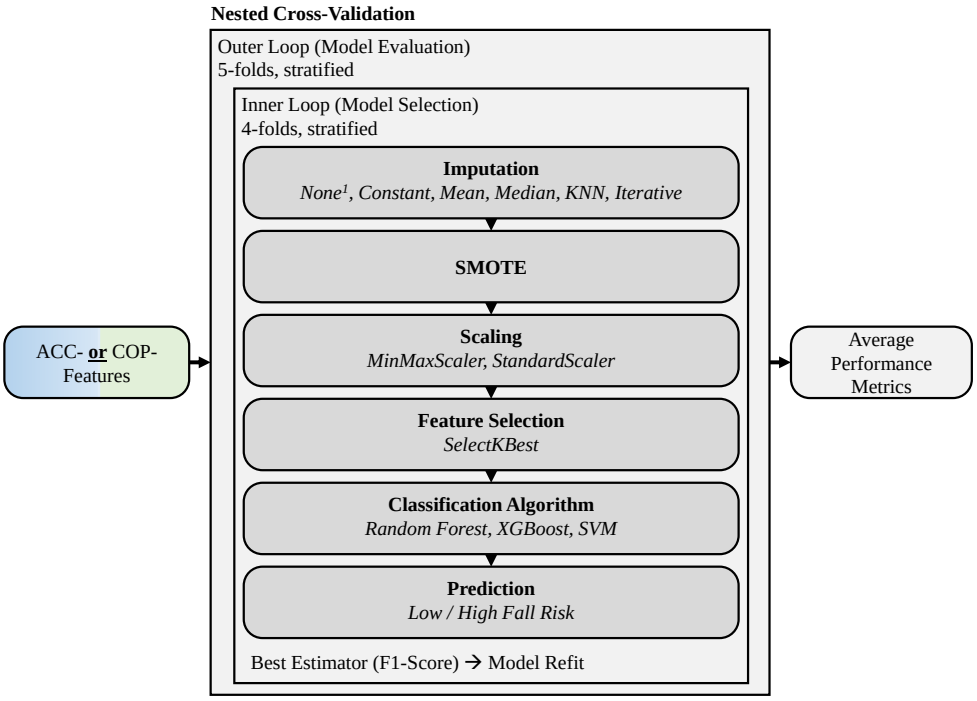


Figure 4.3: Nested cross-validation for fall risk classification with ACC or COP features. ([1] Only for XGBoost.)

included in the hyperparameter space. A detailed summary of all classifier-specific hyperparameter ranges and search spaces is provided in Table 4.4. As a baseline, XGBoost was also evaluated without imputation and SMOTE subsampling to leverage its native support for missing values and assess performance without resampling. The 5x4 nested cross-validation was implemented using *StratifiedKFold* from *scikit-learn*. Hyperparameter optimization was performed using *RandomizedSearchCV* of the *scikit-learn* library with 300 iterations and F1-Score as the scoring metric to select the best pipeline configuration. This configuration was then retrained on the full inner training set and evaluated on the corresponding outer test fold. The pipeline outputs class probabilities, enabling the calculation of Receiver Operating Characteristic - Area under the Curve (ROC-AUC) alongside F1-Score, as described in Section 4.6, to provide a balanced assessment of the model's discriminatory power and to account for both precision and recall.

Force Plate Features

Similar to the approach used for features derived from the IMU, FP data were also utilized to predict FES-I-based fall risk classes. From the FP signals, COP features were extracted to characterize postural sway during static balance tasks. These features offer a well-established biomechanical

representation of balance control and have traditionally served as a gold standard in posturographic analysis. The same machine learning pipeline as for the IMU data, including imputation, SMOTE-based class balancing, scaling, and nested cross-validation with different classifiers, was employed to evaluate the predictive performance of COP-based features. Beyond their role in classification, the predictions based on COP data served as a reference to contextualize and validate the predictive performance of the IMU-based models, with the generalized pipeline presented in Figure 4.3.

4.5 MoCA Classification

Since falls occur twice as frequently in individuals with cognitive impairments, and clinical evidence has demonstrated a close relationship between cognitive function and balance performance in older adults, cognitive function was considered a relevant target for predictive modeling using sway-related features [Mon12]. Accordingly, in addition to fall risk classification using the FES-I score, the same machine learning pipeline was applied to predict cognitive impairment based on the Montreal Cognitive Assessment (MoCA) score. A threshold of 26 was used to differentiate between participants with normal cognitive performance (MoCA ≥ 26, $N = 37$) and those with cognitive impairment (MoCA < 26, $N = 22$) [Nas05]. All feature processing steps, including imputation, SMOTE-based class balancing, scaling, feature selection, and nested cross-validation with hyperparameter optimization were identical to those described for fall risk prediction using ACC-derived features, with hyperparameters shown in Table 4.4. The classification was conducted separately using ACC-derived features and COP-derived features. Model performance was also evaluated using ROC-AUC and F1-Score to ensure comparability with the fall risk classification task.

4.6 Evaluation Metrics

Pearson's Correlation Coefficient r

To assess the correlation between the features derived from the IMU and the FP, the Pearson correlation coefficient was calculated. Given paired data $\{(x_i, y_i)\}_{i=1}^{n}$, Pearson's correlation coefficient r is computed as [Coh88]:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \ \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Here, $\bar{x}$ and $\bar{y}$ denote the means of the x and y values, respectively. This formula normalizes the covariance by the standard deviations, providing a dimensionless measure of the linear association between x and y. The interpretation of the correlation strength follows conventional thresholds summarized in Table 4.2.

Table 4.2: Conventional correlation effect size thresholds [Coh88].

Correlation Level	**Pearson r**
Small (Low Correlation)	$0.10 \leq \|r\| < 0.30$
Medium (Moderate Correlation)	$0.30 \leq \|r\| < 0.50$
Large (High Correlation)	$\|r\| \geq 0.50$

p-Value for Pearson's Correlation (Two-Tailed Test)

After computing Pearson's correlation coefficient r, a two-tailed hypothesis test can assess whether the observed correlation differs significantly from zero. This is done by converting r into a t-statistic using the formula:

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}} \tag{4.20}$$

where n is the number of paired observations. Under the null hypothesis of no correlation ($r = 0$), this statistic follows a t-distribution with $df = n - 2$ degrees of freedom. The corresponding p-value is calculated as:

$$p = 2\left(1 - F_T\left(|t|;\, df\right)\right) \tag{4.21}$$

where F_T is the cumulative distribution function of the t-distribution. The p-value reflects the probability of observing a correlation as extreme as the one calculated, under the assumption of no true correlation [Vir20]. In this thesis, the thresholds for interpreting p-values from the Pearson correlation follow conventional significance levels, as commonly applied in t-tests, shown in Table 4.3.

Table 4.3: Conventional thresholds for statistical significance.

Significance Level	**p-value**
Not Significant	$p \geq 0.05$
Significant	$0.01 \leq p < 0.05$
Very Significant	$0.001 \leq p < 0.01$
Highly Significant	$0.0001 \leq p < 0.001$

Performance Metrics

Model predictions can be evaluated by determining whether each instance was correctly or incorrectly assigned to the positive or negative class. In the context of fall risk prediction, predictions in which patients at high risk of falling were correctly identified were considered True Positive (TP), while cases in which low-risk patients were incorrectly predicted as high-risk patients were referred to as False Positive (FP). Missed detections, where high-risk individuals were classified as low risk, were False Negative (FN), and correct identifications of low-risk patients were True Negative (TN). The performance of a model is typically evaluated using several metrics, with accuracy being one of the most common:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.22}$$

While accuracy is a common metric, it can be misleading in imbalanced datasets. A model may achieve high accuracy by predominantly predicting the majority class, neglecting the minority class's performance. Since minimizing missed detections (FN) is particularly important in fall risk prediction, because failing to identify high-risk individuals may lead to preventable falls, the Recall can be used to evaluate the models. Recall measures the proportion of actual positives that were correctly identified. Meanwhile, the Precision indicates the proportion of positive predictions that are actually correct:

$$\text{Recall} = \frac{TP}{TP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP} \tag{4.23}$$

These two metrics form the basis of the F1-Score, which is the harmonic mean of precision and recall, providing a balance between false positives and false negatives:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.24}$$

While the F1-Score provides a single summary metric based on a specific decision threshold, typically set at 0.5, at which the model classifies an observation as positive or negative, the performance of the model at different thresholds is not captured. To address this, the Receiver Operating Characteristic (ROC) curve can be used, which evaluates the model's ability to distinguish between classes independently of any specific decision threshold. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across various decision thresholds, where each threshold determines the cutoff for classifying predictions as positive or negative based on the

model's output probabilities. The TPR (equivalent to recall) and FPR are computed as:

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN} \tag{4.25}$$

By varying the decision threshold between 0 to 1, the ROC curve illustrates the trade-off between true positive and the false positive rate. The overall performance across all thresholds is summarized by the ROC-AUC). The ROC-AUC can be interpreted as the probability that the model assigns a higher predictive value to a randomly selected positive sample than to a randomly selected negative sample. An AUC of 0.5 indicates performance no better than random chance, while an AUC of 1.0 represents perfect discrimination between classes. In the context of fall risk prediction, a higher ROC-AUC reflects a better ability to correctly identify at-risk individuals across a range of decision thresholds. To evaluate the performance of the classification models, the F1-Score and ROC-AUC can be used, providing more meaningful insights than accuracy, especially for imbalanced two-class problems. In such scenarios, accuracy may overestimate performance by favoring the majority class, while the F1-Score and ROC-AUC offer a balanced evaluation of both positive and negative classes [Sai15].

Table 4.4: Hyperparameter grid used for fall risk classification with IMU- and FP-based features using the *scikit-learn*, *xgboost*, and *imbalanced-learn* libraries.

Imputation Style	**Hyperparameter**	**Values**
SimpleImputer	strategy	mean, median
KNNImputer	n_neighbors	3, 4, 5
Oversampling Technique		
SMOTE	k_neighbors	3, 4, 5
Feature Selection		
SelectKBest	k	20, 50, all
Classifier		
XGBoost	eval_metric	logloss, auc
	learning_rate	0.01, 0.05, 0.1
	n_estimators	50, 100, 250, 500
	max_depth	4, 8, 16, None
	subsample	0.7, 0.8, 1.0
	colsample_bytree	0.7, 0.8, 1.0
Random Forest	bootstrap	True, False
	criterion	gini, entropy
	max_depth	4, 8, 16, 32, None
	max_features	0.1, 0.3, 0.5, sqrt
	min_samples_leaf	0.1, 0.2
	min_samples_split	0.3, 0.4, 0.5, 0.6
	n_estimators	50, 100, 250, 500
	ccp_alpha	0.0, 0.3
SVM	kernel	linear, rbf, poly
	C	0.01, 0.1, 1, 10, 100
	gamma[1]	0.0001, 0.001, 0.01, 0.1, 1, 10
	degree[2]	2, 3, 4, 5, 6

[1] Only for RBF kernel.
[2] Only for polynomial kernel.

Chapter 5

Results and Discussion

In the following chapter, the results obtained with the methods explained in Chapter 4 are presented and discussed in detail. First, the correlation results between features derived from an earable IMU and FPs are analyzed to assess the technical validity of the IMU measurements. This is followed by the evaluation of classification results for fall risk prediction based on the FES-I using IMU- and FP-derived features, including the influence of different imputation strategies and classifiers. Additionally, the results of cognitive function prediction using MoCA-Scores are presented. Each section concludes with a discussion interpreting the findings.

5.1 Correlation between Earable IMU and Force Plate Features

The first aim of this thesis was to technically validate the use of earable IMUs by correlating them with FPs, the gold-standard for balance assessment.

Results

The correlation results are visualised in Figure 5.1 and Figure 5.2, which show heatmaps of correlation strength (r) and associated statistical significance (p) for each of the twelve extracted features across ten static balance conditions. No Bonferroni correction was applied as the analysis focussed on identifying general correlation trends rather than testing isolated hypotheses. Given the exploratory nature of the study and the interdependence of the features, a strict correction for multiple comparisons was not considered appropriate.
Based on the mean correlation strength across tasks, mostly time- and area-domain features, such as DIST, RMS, PATH, MV, RANGE, and AREA, exhibited high correlations (mean $r \geq 0.5$) and MF with medium correlation, as shown in Figure 5.1. In contrast, for the frequency-domain

features, only the PWR feature reached large correlation strength in the mean value, while F95 and CF showed medium correlations, MF and F50 showed small correlations, and FD displayed no meaningful correlation across tasks. Across all tasks, the highest correlation was found for the PWR feature in the EC closed task (r=0.97) and in the AREA feature for the EC wide task (r=0.88). Figure 5.2 indicates that the majority of correlations were statistically significant (p<0.05), and many were highly significant (p<0.001). The exceptions were mostly limited to weaker correlated features in the frequency-domain and more demanding tasks such as EC Tandem or EO Semi.

To contextualize the findings of this work, the results of the EO wide static balance task were compared to those reported in the ISway [Man12] study and presented with correlation strength (r) and the associated statistical significance (p) in Figure 5.3 and Figure 5.4. Both studies evaluated correlations between IMU- and FP-derived features during a static balance task with EO and a wide stance. While this work used an IMU positioned at the ear, the ISway study placed the IMU at the trunk. Across both studies, almost all features showed at least a medium correlation. This work showed a larger correlation than the correlations presented by ISway for several features, including PATH, RANGE, MV, MF, AREA, PWR and CF features, while showing weaker correlations in DIST, RMS, F50, and F95 features. Meanwhile, the FD feature exhibited no correlation and was not statistically significant in the present work, while showing larger correlation results in the ISway study. Besides to the FD feature, the statistical significance of all the features in this study is higher than in the ones reported in the ISway study.

To demonstrate the feature value ranges, units, potential outliers, scatterplots for each feature across all tasks are presented in Figure 5.5 and Figure 5.6. To account for possible outliers in the data and to properly represent the linear correlation, a regression line was fitted using Random Sample Consensus (RANSAC) regression with a residual threshold of 0.3, resulting in a robust estimate of the linear relationship between IMU and FP features.

Discussion

The correlation analysis revealed statistically significant relationships for many features, particularly those related to time-domain and area-based measures. Several of these correlations also showed strong effect sizes, indicating a meaningful linear association between IMU- and FP–derived metrics under certain conditions.

Notably, the correlations were stronger under less demanding balance conditions, particularly in the wide stance with EO, where sensory input was unaffected of potential balance problems of the participant and postural control was more stable. In contrast, harder tasks, such as tandem and pad or tasks involving EC conditions, produced lower correlations, likely due to increased

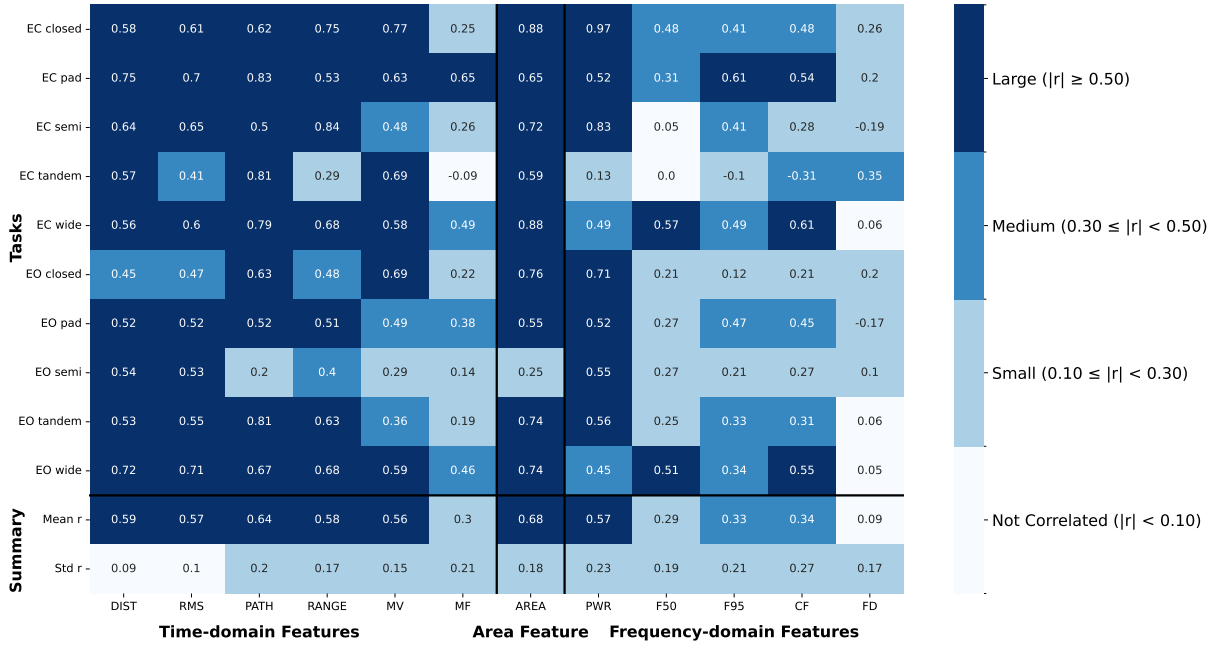


Figure 5.1: Heatmap of Pearson correlation coefficients (r-values) between IMU (ACC) and FP (COP) features across all static balance tasks. Values are rounded to the second decimal place.

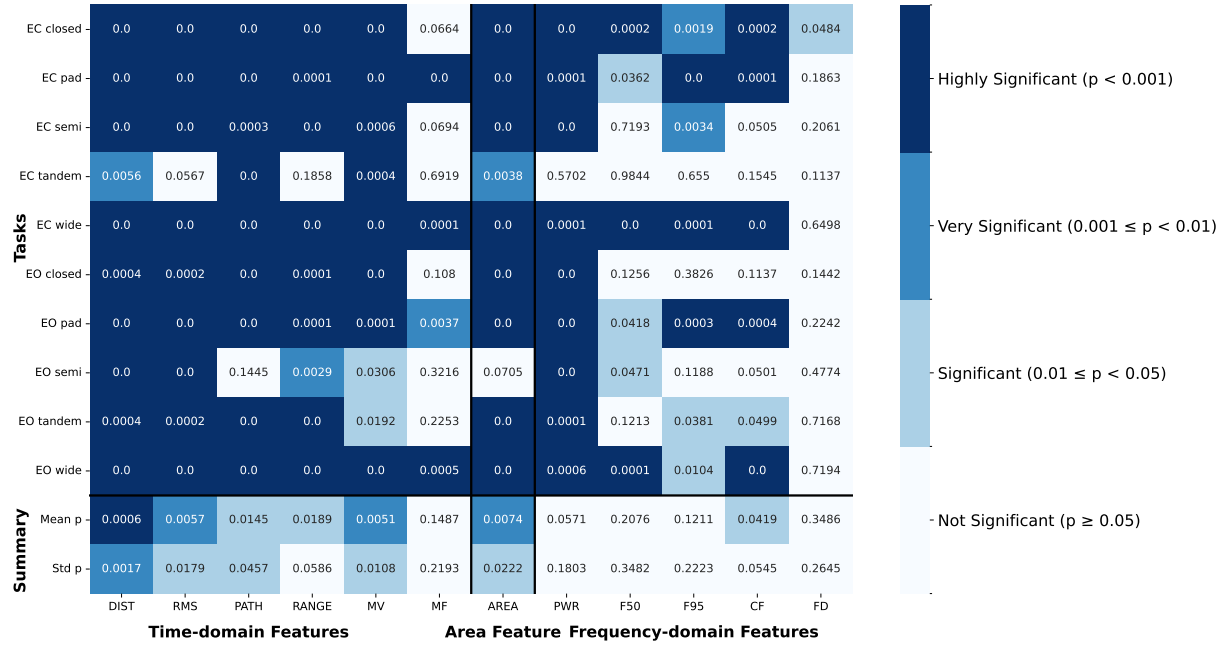


Figure 5.2: Heatmap of statistical significance (p-values) corresponding to Pearson correlations between IMU (ACC) and FP (COP) features across all static balance tasks. Values are rounded to the fourth decimal place.

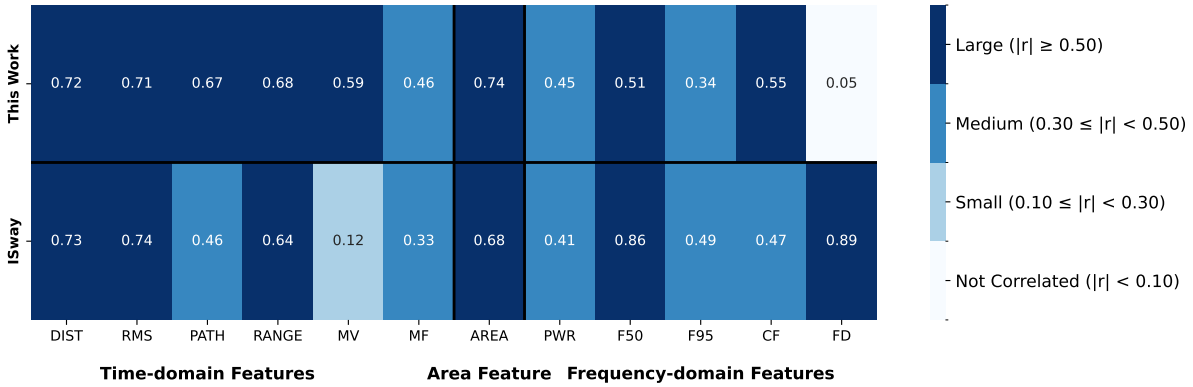


Figure 5.3: Comparison of ACC–COP correlation strengths (r-values, rounded to the second decimal place) between this study and the ISway study [Man12].

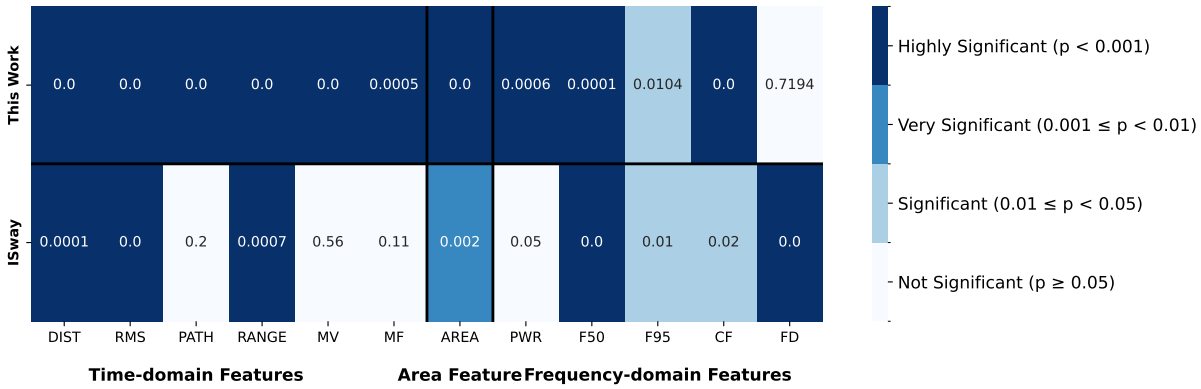
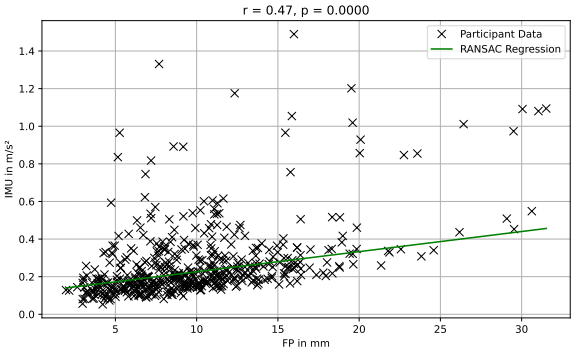
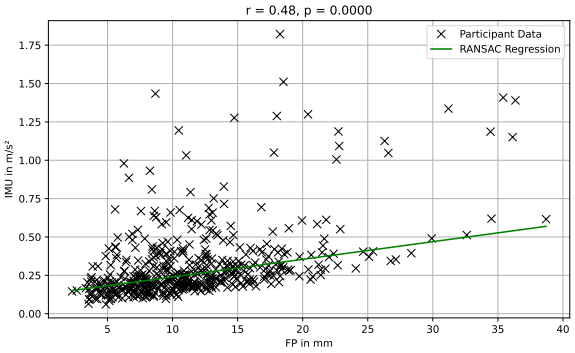


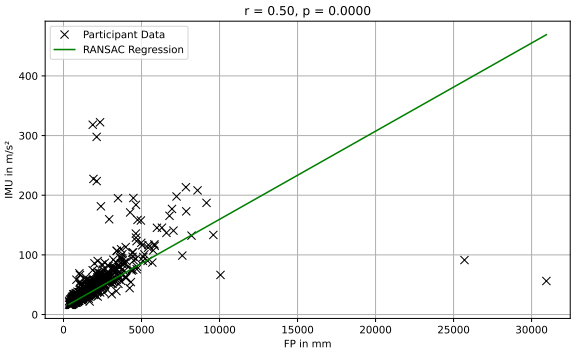
Figure 5.4: Comparison of statistical significance (p-values, rounded to the fourth decimal place) for ACC–COP correlations between this work and the ISway study [Man12].



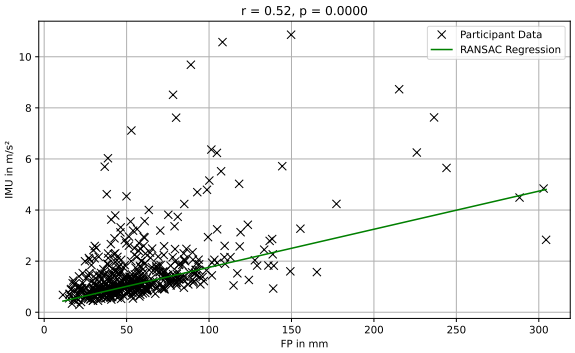
DIST



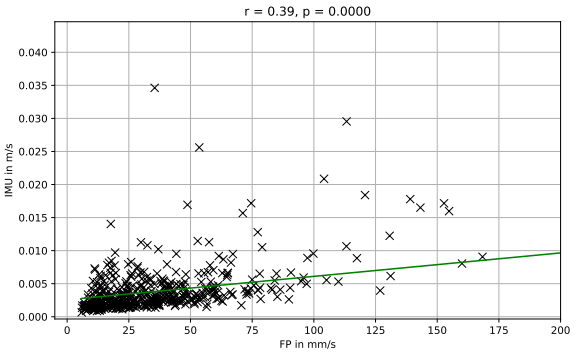
RMS



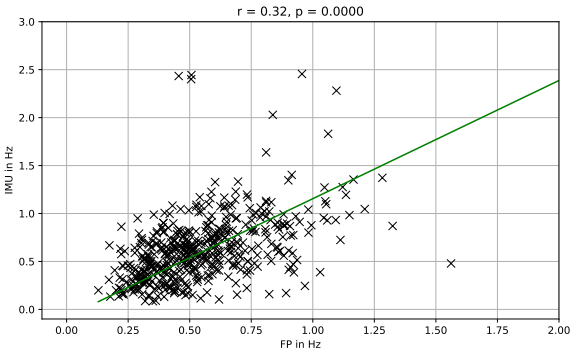
PATH



RANGE

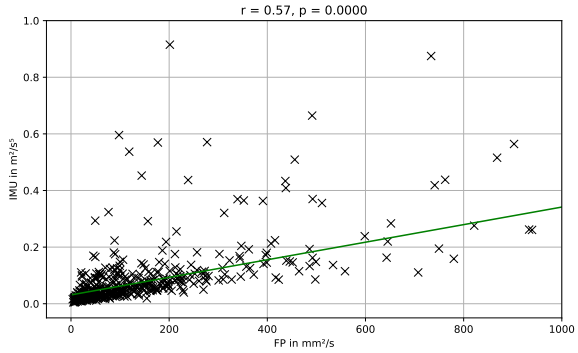


MV (Zoomed)

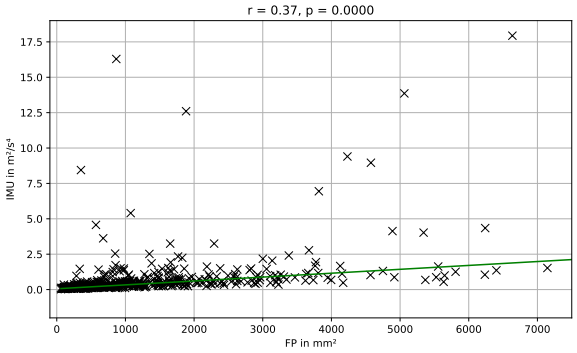


MF (Zoomed)

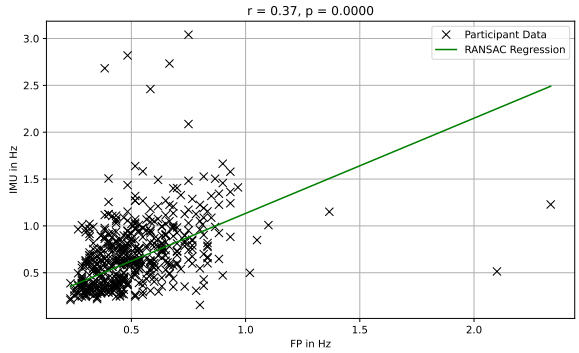
Figure 5.5: Scatterplots illustrating the correlation between IMU- and FP–derived time-domain features aggregated across all standing balance tasks. A regression line fitted using RANSAC is shown for visual reference. Zoomed-in views are shown for MV and MF to improve visual clarity. Full views are provided in the appendix.



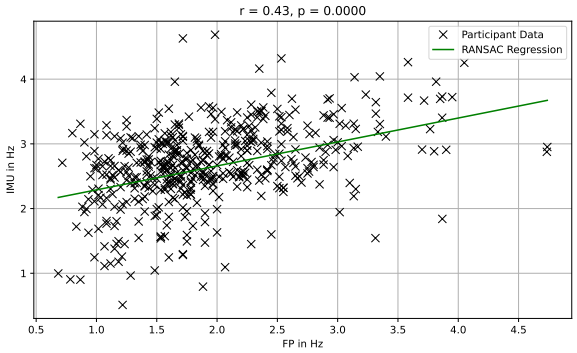
AREA (Zoomed)



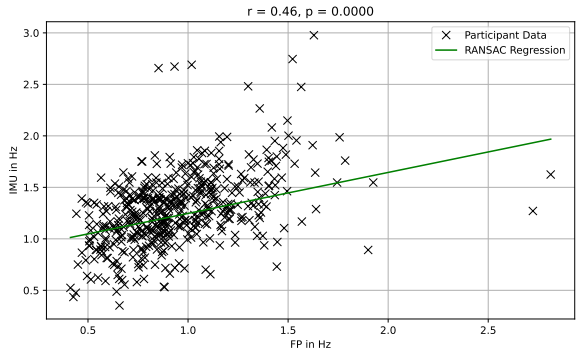
PWR (Zoomed)



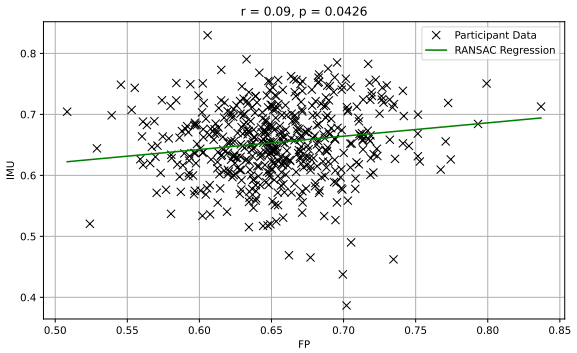
F50



F95



CF



FD

Figure 5.6: Scatterplots illustrating the correlation between IMU- and FP–derived area and frequency-domain features aggregated across all standing balance tasks. A regression line fitted using RANSAC is shown for visual reference. Zoomed-in views are shown for AREA and PWR to improve visual clarity. Full views are provided in the appendix.

movement variability and measurement noise. Furthermore, the strength of the correlations varied across feature domains. Time-domain and area-related features demonstrated higher and more consistent correlations compared to frequency-domain features. This discrepancy may stem from the fact that time and area metrics are directly influenced by gross postural movements, which are similarly captured by both the IMU and the FP. Conversely, frequency-domain features are more sensitive to subtle fluctuations in postural sway and may be affected by anatomical placement. This is particularly relevant for ear-mounted IMU sensors, where sway-related oscillations may be less pronounced than at the center of mass.

The overall strong and statistically significant relationships observed for many features, particularly in the time domain and area-related metrics, support the technical validity of using earable IMUs for postural sway assessment. To contextualize these findings, results from the static balance task in a wide stance with EO were compared to those reported in the ISway study [Man12], which used trunk-mounted IMUs. Although the sensor positions differ substantially, both studies showed similar patterns of correlation strength across sway features, suggesting that key balance-related metrics may be robust to differences in sensor placement. This comparison should nonetheless be interpreted cautiously, as the studies involved different participant groups and protocols. The stronger statistical significance observed in this work is likely due to the larger sample size, which improves the ability to detect true associations, while differences in effect sizes may also be influenced by variation in sensor location or subject characteristics. For this work, 59 participants were included as opposed to only 25 participants in the ISway study. This larger sample increases the likelihood of detecting significant effects due to improved statistical power, which is reflected in the t-statistic formula shown in Section 4.6. The FD feature, which showed the strongest correlation in the ISway study, yielded the weakest correlation in this work. A possible explanation lies in the sensor placement, as data derived from the trunk may capture different aspects of postural control than data collected at the ear.

The scatterplots of the correlations over all tasks reveal some outliers in each feature space, but the majority of participants fall along a linear trend.

These results suggest that, despite some feature-specific differences, earable IMUs can capture sway features that align well with FP measurements. The comparison to previous work using trunk-mounted IMUs further supports the potential of ear-worn sensors for balance assessment.

5.2 Fall Risk Classification

The second and third main objectives of this thesis were to investigate whether sway features obtained from the ACC of a earable IMUs could be used to predict fall risk and to compare this performance to that of COP–derived features of FPs. For this purpose, classification models were trained using both IMU and FP data, with fall risk labels derived from the FES-I score. The classification results are presented separately for ACC- and COP-derived features in the following subsections.

5.2.1 Earable IMU Features

Results

The classification results using IMU-derived features for fall risk prediction are presented in Table 5.1, which reports ROC-AUC and F1-Scores across five outer folds for a fixed random state. Across all classifiers and imputation methods, performance varied notably in both overall level and variance. Random Forest (RF) and Extreme Gradient Boosting (XGBoost) generally outperformed SVM, with RF achieving the highest overall performance when combined with median imputation (ROC-AUC = 0.71 $\pm$0.07; F1 = 0.40 $\pm$0.12). On average across imputations, RF showed the strongest results (mean ROC-AUC = 0.66 ± 0.07), followed closely by XGBoost, while SVM consistently yielded lower classification performance. With respect to imputation methods, median imputation produced the highest mean F1-Score. Mean and k-Nearest Neighbors (KNN) imputation yielded comparable ROC-AUC values but showed more variability across classifiers. Overall, the results indicate rather low predictive capability, showing above-chance performance. The standard deviations reported across folds, especially for XGBoost and SVM, reveal considerable variability in performance.

The classification results using ACC features across multiple classifiers and imputation strategies, averaged over three random states of the outer-fold means from each repetition of the models, are shown in Table 5.2. Overall, performance was relatively consistent across repetitions, with low standard deviations. Tree-based classifiers, namely XGBoost and RF, showed consistently stronger performance than SVM across both ROC-AUC and F1-Score. On average, XGBoost achieved slightly higher ROC-AUC and F1-Scores than RF across imputation methods. Across the different classifiers, the choice of imputation strategy resulted in only minor differences, with ROC-AUC and F1-Scores remaining relatively consistent across Mean, Median, and KNN imputation. All classifiers achieved ROC-AUC values well above 0.5, indicating above-chance performance across configurations.

Table 5.1: Classification performance of IMU-based models for fall risk prediction (random state 12), evaluated across five outer cross-validation folds. Values represent Mean ± standard deviation over folds. "Avg (Mean, Median, KNN)" rows show averaged results across imputation strategies for each classifier. The bottom block reports imputation-wise averages across classifiers. Highest values per metric are highlighted in bold font.

Classifier	**Imputation**	**ROC-AUC**	**F1**
XGBoost	None	0.58 ± 0.11	0.30 ± 0.18
	Mean	0.62 ± 0.13	0.30 ± 0.19
	Median	0.60 ± 0.12	0.37 ± 0.22
	KNN	0.65 ± 0.12	0.33 ± 0.07
	Avg (Mean, Median, KNN)	0.62 ± 0.12	0.33 ± 0.16
Random Forest	Mean	0.63 ± 0.06	0.29 ± 0.18
	Median	**0.71 ± 0.07**	**0.40 ± 0.12**
	KNN	0.64 ± 0.08	0.33 ± 0.10
	Avg (Mean, Median, KNN)	0.66 ± 0.07	0.34 ± 0.13
SVM	Mean	0.64 ± 0.23	0.25 ± 0.15
	Median	0.56 ± 0.10	0.39 ± 0.11
	KNN	0.55 ± 0.26	0.18 ± 0.17
	Avg (Mean, Median, KNN)	0.59 ± 0.20	0.27 ± 0.14
Imputation Method Averages Across Classifiers			
	Mean	0.63 ± 0.14	0.28 ± 0.17
	Median	0.62 ± 0.09	0.39 ± 0.15
	KNN	0.61 ± 0.15	0.28 ± 0.11

To get an insight of the class separation based on ACC sway metrics, Figure 5.7 presents pairplots of the features recorded during one of the most challenging condition, standing with EC on a foam pad. The plots display the joint distributions of feature pairs alongside color-coded class labels (low vs. high fall risk). Across all visualized features, there is substantial overlap between the two fall risk classes. No feature or feature combination shows a clear separation boundary, and both classes appear broadly distributed across the same value ranges. Even in this difficult condition, which theoretically should amplify postural instability in high-risk individuals, class separation remains minimal. This visual impression aligns with the limited F1-Scores observed in the corresponding classification models.

Table 5.2: Overall classification performance of IMU-based models for fall risk prediction, averaged across three random states. For each random state, the mean score over its outer cross-validation folds was computed, and the reported values (mean ± standard deviation) reflect the average and variability of these random state–level means. "Avg (mean, median, KNN)" rows reflect averages across imputation strategies per classifier. Highest values per metric are highlighted in bold font.

Classifier	**Imputation**	**ROC-AUC**	**F1**
XGBoost	None	0.62 ± 0.05	0.30 ± 0.08
	Mean	0.63 ± 0.05	0.36 ± 0.05
	Median	0.63 ± 0.00	**0.37 ± 0.02**
	KNN	**0.66 ± 0.08**	0.33 ± 0.02
	Avg (Mean, Median, KNN)	0.64 ± 0.04	0.35 ± 0.02
Random Forest	Mean	0.64 ± 0.03	0.33 ± 0.07
	Median	0.63 ± 0.08	0.37 ± 0.06
	KNN	0.63 ± 0.02	0.31 ± 0.06
	Avg (Mean, Median, KNN)	0.63 ± 0.03	0.34 ± 0.03
SVM	Mean	0.54 ± 0.01	0.34 ± 0.07
	Median	0.52 ± 0.08	0.33 ± 0.04
	KNN	0.53 ± 0.01	0.31 ± 0.08
	Avg (Mean, Median, KNN)	0.53 ± 0.02	0.33 ± 0.02
Imputation Method Averages Across Classifiers			
	Mean	0.60 ± 0.06	0.34 ± 0.06
	Median	0.59 ± 0.08	0.36 ± 0.04
	KNN	0.61 ± 0.06	0.32 ± 0.06

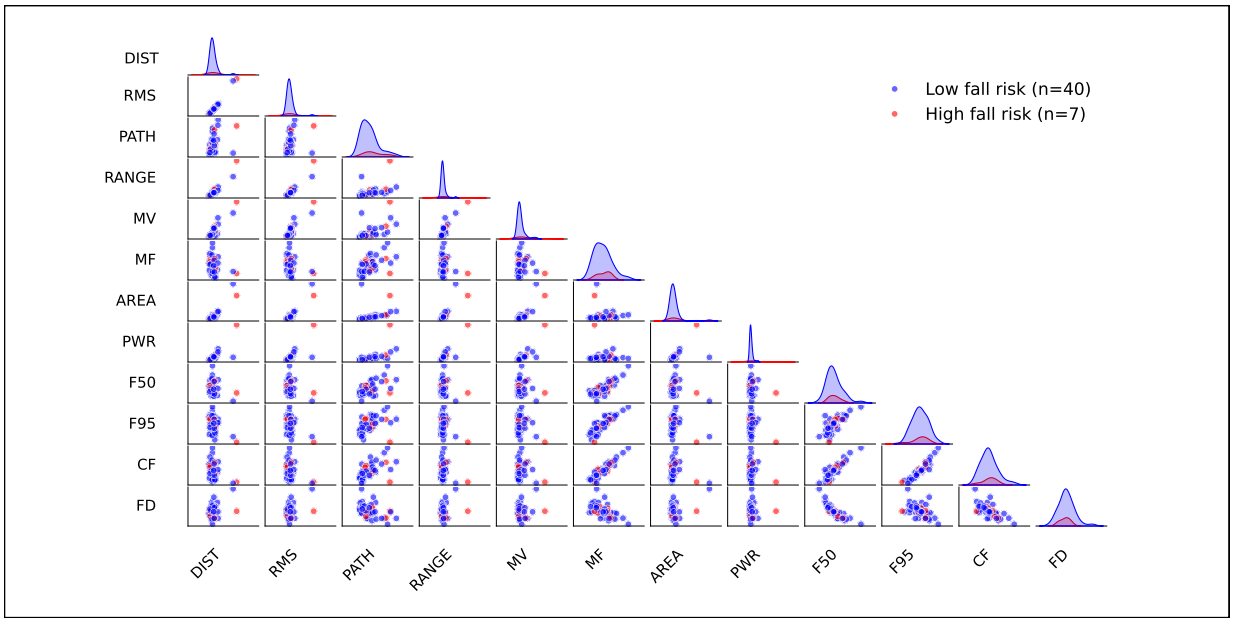


Figure 5.7: Example pairplot of ACC sway metrics during standing task on a pad with EC for low fall risk and high fall risk groups. Additional selected pairplots of other difficult tasks are provided in the appendix.

Discussion

The evaluation of model performance across five outer folds using a fixed random state revealed considerable variability in both ROC-AUC and F1-Scores, particularly for XGBoost and SVM classifiers. This fold-to-fold variation reflects poor generalization and suggests that the models may be overly sensitive to how training and validation data are partitioned. Standard deviations of up to ±0.26 for ROC-AUC and ±0.22 for F1-Score indicate that classifier performance is not consistent across subsets of the data, undermining confidence in the stability of the prediction of FES-I classes. While moderate classification performance was achieved in some configurations, the low F1-Scores and high performance variance show the need for larger, more balanced datasets.

Overall, across different random states, the classification performance of IMU-based models remained relatively consistent, with only minor variations in ROC-AUC and F1-Scores. The low standard deviations suggest that, regardless of the random initialization, the models tended to classify the data in a similar way. This consistency across random states may reflect a lack of distinct, class-specific patterns in the feature space. Rather than discovering fundamentally different solutions, the models appear to converge on similar decision boundaries, not because they generalize well, but because the available features offer only limited discriminative information. The low variability in performance suggests that, regardless of the initialization, the models are

consistently constrained by the same weak information, leading to stable but modest classification outcomes.

These limitations are further reflected in the similar classification outcomes observed across different imputation strategies. Whether missing task feature data were handled using Mean, Median, or KNN imputation, the resulting ROC-AUC and F1-Scores varied only slightly across classifiers. This uniformity suggests that the predictive information is not strongly tied to a specific method of handling missing values, likely because the underlying feature space lacks sharply defined boundaries between classes. While median imputation yielded marginally higher F1-Scores on average, the overall small differences imply that no imputation strategy is able to recover substantial additional class-relevant information from the incomplete data.

The pairplot analysis further supports the quantitative findings by offering a qualitative view of the underlying feature space. Despite the use of a challenging condition with EC on foam pad, there is no apparent visual separation between low- and high-risk groups in any of the sway features. From a clinical perspective, such difficult tasks are expected to magnify balance deficits and thus improve class discriminability. However, this anticipated pattern is not evident in the data. Consequently, the models classify based on features that are largely overlapping between classes, limiting predictive performance.

The absence of clear class separation in both visualizations and model outputs suggests that the ACC-derived sway features, as used in this study, did not sufficiently differentiate between fall risk groups. This limitation may be due to two possible factors. On one hand, the ear-worn IMU may capture less detailed or discriminative information about postural control compared to more established measurement systems. On the other hand, the dataset may not provide enough variability or contrast in fall risk to support the development of reliable classification models. This first possibility will be further explored in the next section, where the same classification pipeline is applied to FP–derived features. Although the ACC sway features of the IMU showed good correlation with COP measurements of the FP, they may not capture differences that clearly distinguish between fall risk groups. Alternatively, the dataset may be limited by factors such as small sample size, imbalance between classes, potential discrepancy between self-reported fall risk based on the FES-I and actual postural control, or the use of static balance tasks that may not evoke sufficiently strong postural responses to reveal subtle impairments. These constraints could mask meaningful group differences and reduce the effectiveness of any model trained on the data. If the classification performance remains similarly low when using COP features of the FP, this would support the interpretation that the main limitation lies in the dataset rather than in the sensing modality.

Previous studies have investigated fall risk classification using IMUs, often achieving stronger results than those observed in this thesis. For instance, Weiss et al. [Wei13] combined performance-based clinical scores with features extracted from 3-day trunk-mounted IMU recordings in older adults. They showed that incorporating free-living acceleration metrics, such as total activity duration and AP acceleration range, substantially enhanced the model's accuracy to identify future fallers to 94.7%, with improved sensitivity and perfect specificity. This demonstrates how real-world, dynamic gait patterns can provide valuable insight into fall risk that may not be captured through brief clinical assessments alone. In contrast, this thesis focused on static postural sway during brief standing tasks, which may be less effective in revealing subtle impairments in this population. Similarly, Del Din et al. [Del19] achieved good faller discrimination using ACC data from multiple sensor locations during walking, suggesting that dynamic assessments better capture context-specific instabilities than static tasks.

5.2.2 Force Plate Features

Results

The classification results using COP-derived features from the FP for fall risk prediction are presented in Table 5.3, which reports ROC-AUC and F1-Scores across five outer folds using a fixed random state. Among all configurations, RF achieved the strongest results, particularly when paired with Mean imputation (ROC-AUC = 0.75 ± 0.19; F1 = 0.59 ± 0.21). SVM also showed competitive performance, when paired with KNN imputation. On average, the different imputation strategies, mean, median, and KNN, yielded comparably similar results, with only minor differences in ROC-AUC and F1-Scores across classifiers. On average across imputations, RF outperformed all other classifiers, followed by SVM, and then XGBoost, for this random state evaluation. The standard deviations reported across folds reflect similar high fold-to-fold variability as observed in the IMU models.

The classification results using COP-derived features, averaged across three random states, are shown in Table 5.4. Overall, the performance is consistent across repetitions, with low standard deviations indicating stable evaluation across different data splits. Among the classifiers, RF consistently achieved the best performance, with the highest ROC-AUC value of 0.71 ± 0.08 and F1-Score of 0.47 ± 0.08 using Mean imputation. On average across imputations, RF yielded the highest scores, followed by XGBoost and SVM. SVM showed more variable performance than the tree-based models, with F1-Scores ranging from 0.34 to 0.42 and higher standard deviation values, depending on the imputation strategy. The imputation methods performed comparably

Table 5.3: Classification performance of FP-based models for fall risk prediction (random state 12), evaluated across five outer cross-validation folds. Values represent mean ± standard deviation over folds. "Avg (Mean, Median, KNN)" rows show averaged results across imputation strategies for each classifier. The bottom block reports imputation-wise averages across classifiers. Highest values per metric are highlighted in bold font.

Classifier	**Imputation**	**ROC-AUC**	**F1**
XGBoost	None	0.58 ± 0.17	0.37 ± 0.25
	Mean	0.61 ± 0.12	0.39 ± 0.12
	Median	0.56 ± 0.20	0.33 ± 0.25
	KNN	0.61 ± 0.11	0.42 ± 0.12
	Avg (Mean, Median, KNN)	0.60 ± 0.14	0.38 ± 0.16
Random Forest	Mean	**0.75 ± 0.19**	**0.59 ± 0.21**
	Median	0.70 ± 0.27	0.45 ± 0.37
	KNN	0.72 ± 0.21	0.50 ± 0.22
	Avg (Mean, Median, KNN)	0.72 ± 0.22	0.51 ± 0.27
SVM	Mean	0.65 ± 0.28	0.48 ± 0.23
	Median	0.59 ± 0.20	0.36 ± 0.29
	KNN	0.71 ± 0.20	0.49 ± 0.21
	Avg (Mean, Median, KNN)	0.65 ± 0.23	0.44 ± 0.24
Imputation Method Averages Across Classifiers			
	Mean	0.67 ± 0.20	0.49 ± 0.19
	Median	0.62 ± 0.22	0.38 ± 0.30
	KNN	0.68 ± 0.17	0.47 ± 0.18

overall, with only minor variations in both ROC-AUC and F1-Scores across classifiers.

To further assess the class separation capability of COP-derived sway features, Figure 5.8 presents pairplots for the features in the EC foam pad condition. This scenario represents one of the most challenging balance tasks and was chosen to highlight potential postural differences between low and high fall risk individuals. The plots illustrate that, despite the increased task difficulty, there is considerable overlap between the two classes across all feature combinations. No clear visual boundary or clustering is apparent, and the feature distributions show a high degree of similarity between the groups. This aligns with the overall classification results, where F1-Scores remained moderate even under high-performing configurations.

Table 5.4: Overall classification performance of FP-based models for fall risk prediction, averaged across three random states. For each random state, the mean score over its outer cross-validation folds was computed, and the reported values (mean ± standard deviation) reflect the average and variability of these random state–level means. "Avg (mean, median, KNN)" rows reflect averages across imputation strategies per classifier. Highest values per metric are highlighted in bold font.

Classifier	**Imputation**	**ROC-AUC**	**F1**
XGBoost	None	0.64 ± 0.03	0.35 ± 0.07
	Mean	0.65 ± 0.02	0.37 ± 0.03
	Median	0.67 ± 0.06	0.41 ± 0.12
	KNN	0.64 ± 0.04	0.39 ± 0.07
	Avg (Mean, Median, KNN)	0.65 ± 0.04	0.39 ± 0.04
Random Forest	Mean	**0.71 ± 0.08**	**0.47 ± 0.08**
	Median	0.67 ± 0.09	0.45 ± 0.08
	KNN	0.69 ± 0.04	0.46 ± 0.02
	Avg (Mean, Median, KNN)	0.69 ± 0.05	0.46 ± 0.03
SVM	Mean	0.64 ± 0.08	0.37 ± 0.18
	Median	0.61 ± 0.11	0.34 ± 0.13
	KNN	0.66 ± 0.03	0.42 ± 0.10
	Avg (Mean, Median, KNN)	0.64 ± 0.05	0.38 ± 0.07
Imputation Method Averages Across Classifiers			
	Mean	0.67 ± 0.04	0.40 ± 0.11
	Median	0.65 ± 0.06	0.40 ± 0.12
	KNN	0.66 ± 0.03	0.42 ± 0.05

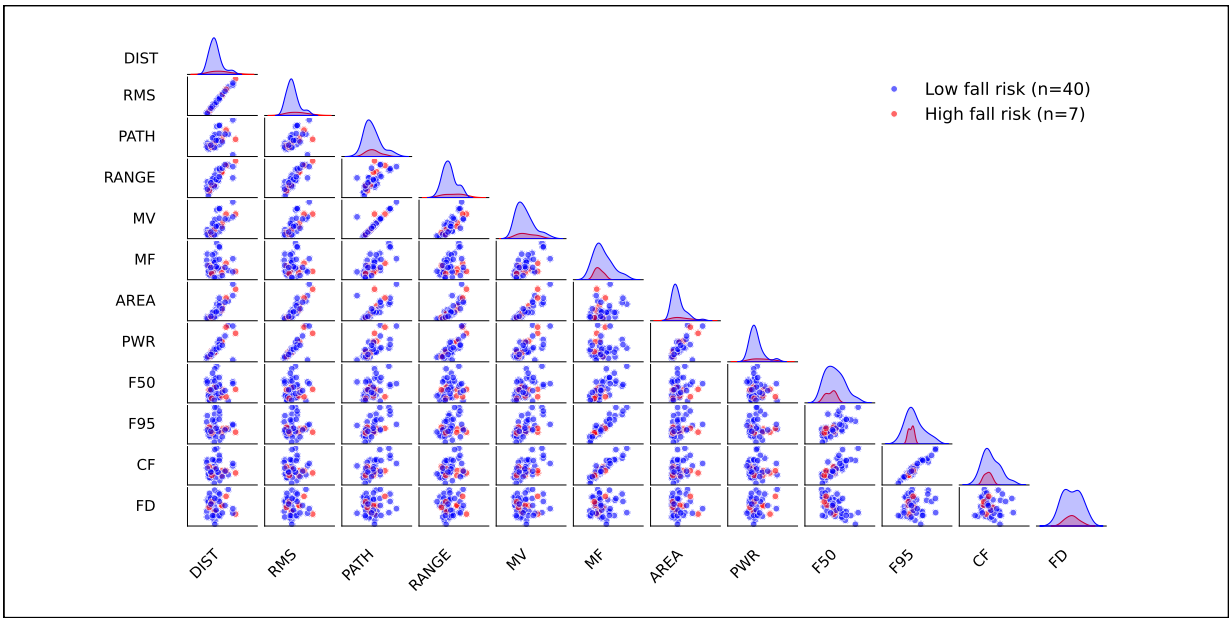


Figure 5.8: Example pairplot of COP sway metrics during standing task on a pad with EC for low fall risk and high fall risk groups. Additional pairplots of other selected difficult tasks are provided in the appendix.

Discussion

Compared to the results obtained using ACC-derived features, the classification models trained on COP-derived features achieved generally higher predictive performance. Both ROC-AUC and F1-Scores were consistently higher across classifiers and imputation methods, with several configurations exceeding 0.7 for ROC-AUC and 0.5 for F1-Score. This indicates that COP sway features, derived from a more stable and precise measurement system, contain more discriminative information for distinguishing between low- and high-risk individuals. While RF again emerged as the strongest classifier, SVM performed notably better with COP features than with ACC-derived ones, suggesting that the structure of the COP feature space may be more favorable for margin-based classification. Imputation strategies showed only minor influence on model performance, with KNN and Mean imputation yielding slightly more stable results.

The evaluation of FP-based models across three random states showed overall higher and more consistent classification performance compared to IMU-based models. While RF achieved the best average results, both XGBoost and SVM also performed competitively in certain configurations. Despite this improvement, variability across cross-validation folds was still present, particularly in F1-Scores. This suggests that while COP features may capture more discriminative information than ACC-derived sway metrics, but the distinction between fall risk groups remains subtle.

Furthermore, the comparably small impact of different imputation methods indicates that model performance is likely driven more by the underlying information in the features than by the method used to handle missing data.

The classification results based on COP-derived features largely confirmed the patterns observed with ACC-derived features. Although FP models achieved slightly higher ROC-AUC and F1-Scores in some configurations, especially with RF classifiers, no clear separation between fall risk groups was evident in the pairplots or classifier outputs. This suggests that the modest improvement in performance is not due to superior discriminative information in the COP features, but rather reflects shared limitations in the underlying data. The continued lack of robust class separation and high variability across cross-validation folds point to fundamental constraints imposed by the dataset itself. These include the small sample size, class imbalance, and the use of the FES-I as a subjective outcome measure. Moreover, the static balance tasks may not have been sufficiently challenging to reveal subtle postural control differences. The observation that both ear-worn IMUs and FPs resulted in similarly limited classification performance suggests that the key limitations are more likely related to the labeling approach and assessment protocol than to the modalities themselves.

The moderate classification results obtained with FP–derived features align with earlier findings on the use of postural sway to assess fall risk. In a comprehensive review, Piirtola and Era [Pii06] concluded that although some FP parameters, particularly those related to ML sway, were associated with future falls, the predictive utility of such measures in isolation was limited. They emphasized inconsistency across studies and the lack of prospective data linking sway to actual fall outcomes. In a more recent study, Sun et al. [Sun19] used machine learning to classify fall risk in individuals with multiple sclerosis based on static FP sway data and reported high accuracy, achieving over 86% classification accuracy using RF on static tasks. These results suggest that under certain clinical conditions and with well-defined task settings, static sway features can yield strong predictive performance. Unlike the work of Sun et al., which derived fall risk labels from objective physiological assessments using the Physiological Profile Assessment (PPA), the present study used a more heterogeneous population and subjective fall risk labels (FES-I). This likely contributed to the lower classification scores observed here. The overall consistency in moderate outcomes across both IMU and FP models further supports the interpretation that task design, population characteristics, and label quality play a greater role in limiting model performance than sensor modality alone.

5.3 MoCA Classification

An additional aim of this thesis was to extend the evaluation of sway-based features beyond fall risk classification and assess their predictive value for other clinically relevant outcomes, such as cognitive function. To this end, classification models were trained to differentiate cognitive status using MoCA scores. The same machine learning pipeline applied in the fall risk classification was applied, incorporating both earable IMU and FP–derived features separately, thereby enabling a direct comparison between fall risk and cognitive classification performance.

Results

The classification results for cognitive status prediction based on MoCA scores are summarized in Table 5.5, reporting model performance across three random states. Both IMU- and FP-based models yielded moderate results across classifiers and imputation strategies, with ROC-AUC values reaching up to 0.64 and F1-Scores up to 0.52.
For ACC-derived features of the IMU, XGBoost with KNN imputation produced the highest F1-Score (0.51 ± 0.01), with similarly strong results observed across other classifiers and imputations. FP models trained on COP features slightly outperformed their IMU counterparts in most configurations. Notably, XGBoost with KNN also achieved the highest ROC-AUC (0.64 ± 0.02) and F1-Score (0.52 ± 0.07) among the FP models.
Overall, classification performance was relatively stable across random states, with low to moderate standard deviations. Both sensor modalities showed classification ability above chance, with COP features offering a modest advantage in discriminative power. Differences between imputation strategies remained minor, consistent with findings from the fall risk prediction task.

Discussion

The classification of cognitive status based on sway features resulted in moderate predictive performance across both IMU and COP modalities. Compared to the fall risk classification task, the MoCA classification yielded similar or slightly better F1-Scores, particularly for the IMU-based models. This may be partially explained by the more balanced distribution of classes in the MoCA dataset, which reduces the impact of class imbalance on model learning and evaluation. And the MoCA might contain more objective information as a target variable compared to the more subjective nature of the FES-I questionnaire, which may have an impact on the sway characteristics. Although COP features again outperformed ACC features in most configurations, the difference was less pronounced than in the fall risk classification task. In both modalities, the highest-performing

Table 5.5: Classification performance of ACC- and COP-based models for MoCA prediction, averaged across three random states. Values represent the mean ± standard deviation of outer-fold scores computed separately for each random state and then aggregated. The values in bold show the best results for ACC and COP.

Classifier	Imputation	ROC-AUC		F1-Score	
		ACC	COP	ACC	COP
XGBoost	None	0.58 ± 0.04	0.63 ± 0.03	0.44 ± 0.02	0.52 ± 0.04
	Mean	0.56 ± 0.02	0.55 ± 0.06	0.50 ± 0.01	0.45 ± 0.04
	Median	0.58 ± 0.05	0.55 ± 0.08	0.48 ± 0.03	0.50 ± 0.05
	KNN	0.57 ± 0.02	**0.64 ± 0.02**	**0.51 ± 0.01**	**0.52 ± 0.07**
Random Forest	Mean	0.60 ± 0.05	0.62 ± 0.05	0.48 ± 0.04	0.49 ± 0.12
	Median	**0.60 ± 0.02**	0.60 ± 0.04	0.44 ± 0.14	0.47 ± 0.06
	KNN	0.59 ± 0.04	0.58 ± 0.09	0.49 ± 0.07	0.46 ± 0.05
SVM	Mean	0.58 ± 0.04	0.60 ± 0.08	0.47 ± 0.08	0.44 ± 0.03
	Median	0.54 ± 0.02	0.60 ± 0.02	0.47 ± 0.05	0.47 ± 0.01
	KNN	0.47 ± 0.04	0.53 ± 0.05	0.42 ± 0.06	0.50 ± 0.04

models achieved ROC-AUCs around 0.64 and F1-Scores slightly above 0.50, suggesting that sway-related features encode some discriminative information related to cognitive status, but with limited strength.

The evaluation behavior of MoCA classification models closely resembled that of fall risk prediction. While the performance remained consistent across random states, the high variability across cross-validation folds persisted, although not shown in detail here. Slightly better results were observed, likely due to a more balanced class distribution and a larger number of samples in the impaired group, but the limited separability in the feature space continued to constrain classification performance. Furthermore, the lack of sharp class separation, suggests that sway metrics alone may not fully capture cognitive impairment. The results of this thesis indicate that while an association between cognitive status and sway features is plausible, as cognitive decline may affect balance, ACC-derived sway metrics are likely to capture only indirect signs of cognitive impairment and do not measure it directly. Therefore, combining sway features with other physiological or behavioral markers might be necessary to build more robust models for cognitive screening.

Only limited work has specifically examined the relationship between sway features and MoCA scores. Apthorp et al. [Apt20] investigated this link in individuals with Parkinson's disease and found that greater postural sway, measured via FPs, was significantly correlated with lower

MoCA scores and reduced quality of life. This suggests that cognitive decline may manifest as increased instability, even in quiet stance. In a related but broader context, Mirelman et al. [Mir12] assessed executive function in older adults and demonstrated that cognitive impairments were predictive of future falls, particularly under dual-task walking conditions. While they did not use MoCA specifically, their findings highlight the interdependence of cognitive function and postural control. In contrast to these studies, the present work used static sway tasks and ear-worn IMUs to predict MoCA-defined cognitive impairment, yielding only moderate classification results. These discrepancies likely reflect differences in sensor modality, task design, and population. They also emphasize that while sway features may indirectly reflect cognitive status, more complex or multimodal assessments are likely needed for accurate classification.

Chapter 6

Discussion and Limitations

6.1 General Discussion

This thesis investigated whether postural sway features derived from the ACC of an ear-worn IMU and COP features derived from FPs could be used to assess fall risk and cognitive status using traditional machine learning models. The findings demonstrate that both sensor modalities provide relevant information, but differ in their discriminative strength and clinical applicability.

The first goal of this thesis was to assess how well sway features captured by ear-worn IMUs correlate with FP metrics, the clinical gold standard for balance assessment. The correlation analysis revealed strong and statistically significant relationships, particularly for time-domain and area-based features, under stable conditions such as standing with EO in a wide stance. Notably, when compared to the ISway [Man12] study, which used a trunk-mounted IMU, this work showed that ear-worn sensors achieved equal or even stronger correlations for several features, despite the greater anatomical distance from the center of mass. These results support the technical feasibility of earable IMUs for use in balance analysis, especially given their portability and unobtrusiveness.

The second research objective was to evaluate the extent to which ear-worn IMU data could be used to predict fall risk using supervised machine learning classifiers. For all models and imputation strategies, the IMUs-based models with ACC-derived features achieved modest but above-chance performance, with limited sensitivity in correctly identifying individuals at high risk of falling (ROC up to 0.71 ± 0.07; F1 up to 0.40 ± 0.12). Performance across different random states remained relatively consistent, indicating that the models converged on similar decision boundaries regardless of initialization, likely due to the limited complexity of the feature space and limited sample diversity. The consistent performance likely indicates that the feature space contains only limited structure, offering few distinct patterns that the models can learn. This is

likely due to overlapping class distributions and a lack of strongly discriminative information in the available features. This is supported by the high variability across outer cross-validation folds, revealing poor generalizability and indicated that the classifiers were strongly influenced by the specific partitioning of the data.

The third research objective, predicting fall risk using FP-based machine learning models, resulted in an overall slightly stronger performance. Nonetheless, performance variability across folds persisted, and class overlap remained also evident in the COP-derived feature distributions. This highlights that although FP data provides a more robust signal, postural sway features alone are not sufficient for highly accurate fall risk prediction based on FES-I target labels.

An additional aim of this thesis was to assess whether sway features could predict cognitive function, using the MoCA score as a target. Overall classification performance was similar to or slightly better than in the fall risk prediction. This was especially true for IMU-based models, which probably benefited from a more balanced class distribution. The moderate scores achieved suggest that sway features contain some indirect information of cognitive impairment, possibly reflecting increased variability or instability in physical control. However, as with fall risk, class overlap in the feature space limited the performance outcome. Additional features could improve prediction accuracy in cognitive screening, for instance, combining sway data with speech and gait data.

Across both IMU and FP modalities, this thesis found only moderate classification performance for fall risk and cognitive status, in contrast to some earlier studies reporting higher accuracies. This discrepancy can be attributed to several key factors. First, many studies, such as those by Weiss et al. [Wei13] and Del Din et al. [del16], used dynamic gait tasks and longer monitoring periods, allowing for richer movement patterns and higher ecological validity. Second, the use of prospective fall labels in studies like FARSEEING [Kle16] and FallRiskPD [Ull22] provides a more accurate ground truth than the retrospective and subjective FES-I scale. The results of this thesis suggest that the predictive ceiling of static postural sway features, particularly under low-challenge conditions, may already have been reached. The inability to reproduce stronger correlations between sway features and clinical outcomes, as reported in older or higher-powered studies, underscores the need for larger and more diverse datasets, real-world assessment protocols, and potentially multi-modal feature sets.

Overall, fall risk prediction using sway features from both earable IMUs and FPs did not yield clinically reliable results. While FP models showed slightly stronger performance, neither modality was able to clearly separate low- and high-risk individuals. This pattern suggests that the limited classification performance may be more strongly influenced by the dataset and study design, such

as the use of a subjective outcome label (FES-I) and the relatively low challenge of the static balance tasks, than by the sensing modality itself. Taken together, these findings suggest that earable IMUs appear technically suitable for capturing postural sway, but their effectiveness for fall risk prediction depends strongly on the task design, outcome label, and overall study context.

6.2 Limitations

While this thesis offers valuable insights into the potential of sway-based features for clinical classification, several limitations must be acknowledged.

First, the fall risk classification in this study was based on the FES-I questionnaire, which captures perceived rather than actual fall risk. Consequently, the models were trained to predict self-reported concern about falling rather than confirmed fall events, which limits the external validity of the results and reduces their applicability to real-world fall prediction scenarios. The FES-I is known to be influenced by psychological and social factors and may not consistently align with physiological indicators of balance impairment. From this perspective, the FES-I might be more suitable as an additional or secondary feature rather than as a primary classification target.

To more accurately predict fall risk from balance-related data, a more appropriate outcome variable that avoids subjective bias and preserves the relationship between sway characteristics and fall-related outcomes would be required. For instance, the FARSEEING project [Kle16] collected real-world IMU data linked to prospectively monitored fall events, enabling the identification of actual fallers instead of relying on retrospective or fear-based questionnaires. Similarly, the FallRiskPD study [Ull22] emphasized the importance of using prospective event-based labeling to differentiate true fallers from those who simply report fear of falling. In contrast, the reliance on FES-I in this thesis may have attenuated the relationship between sway patterns and the classification labels, making it more difficult for the models to extract reliable discriminative features. Additionally, the relatively small sample size and class imbalance, particularly in the high-risk and cognitively impaired groups, restricted the statistical power of the models and contributed to performance variability across cross-validation folds.

Another important limitation stems from missing data. Several participants were unable to complete more challenging balance tasks, resulting in missing features that had to be imputed. Although multiple imputation methods (mean, median, KNN) were tested, and more techniques such as iterative and constant-value imputation were briefly explored, but left out due to significant runtime increase, all imputations introduce some degree of uncertainty.

Another limitation of this study is that it focused exclusively on static balance tasks, which may

not fully capture the complexity of postural control in daily life. Incorporating dynamic tasks such as walking, turning, or dual-task conditions could provide more informative and discriminative features for assessing both fall risk and cognitive function.

A limitation of the correlation approach might be the lack of precise synchronization between the IMU and FP recordings, which may have led to temporal misalignments and reduced the accuracy of the extracted features used for correlation analysis. Future studies should apply synchronization techniques, such as jump-based alignment [Sei12], to improve temporal alignment between sensor signals and enhance the validity of cross-modal comparisons.

The generalizability of the findings is limited by the specific study population, which consisted of older adults, including some with Parkinson's disease. Although this cohort is clinically relevant, the results may not be generalizable to other populations, such as patients at different stages of disease or patients with other medical conditions.

Chapter 7

Conclusion and Outlook

The primary aim of this thesis was to investigate whether balance characteristics derived from a single ear-worn IMU can be used to predict fall risk in older adults. To achieve this, a technical validation was first performed to assess the correlation between sway features extracted from the IMU's ACC and features derived from the COP obtained from simultaneously recorded laboratory FPs, which served as the biomechanical gold standard. The same set of derived sway features from both modalities, including time-domain, area-based, and frequency-domain features, was then used as input to a machine learning pipeline for fall risk classification. The risk of falling was defined as a binary classification target based on the FES-I. The predictive performance of the IMU-based approach was directly compared with that of the FP-based model using the same evaluation framework. All analyses were performed using a pre-recorded dataset that included an earable IMU and FP data collected during static balance tasks.

The outcomes of the technical validation were promising, showing that sway features derived from the ACC of the ear-worn IMU exhibited moderate to strong correlations with corresponding COP features from the FPs. In particular, time-domain features, e.g. PATH and RMS of the trajectory, demonstrated strong correlation between the two modalities. Frequency-domain features showed moderate correlations, indicating some differences in how dynamic balance aspects are captured. Overall, these results suggest that the IMU captures key aspects of postural control during quiet standing and provides balance-related information that is comparable to that obtained from laboratory-grade FPs. This highlights the potential of ear-worn IMUs as a practical, unobtrusive alternative for assessing balance in clinical settings.

In contrast, classification performance in predicting fall risk using machine learning, employing models such as RF, XGBoost, and SVMs, was moderate for both IMU- and FP–derived features. The models showed limited discriminatory power with ROC-AUC values of up to 0.71 and F1

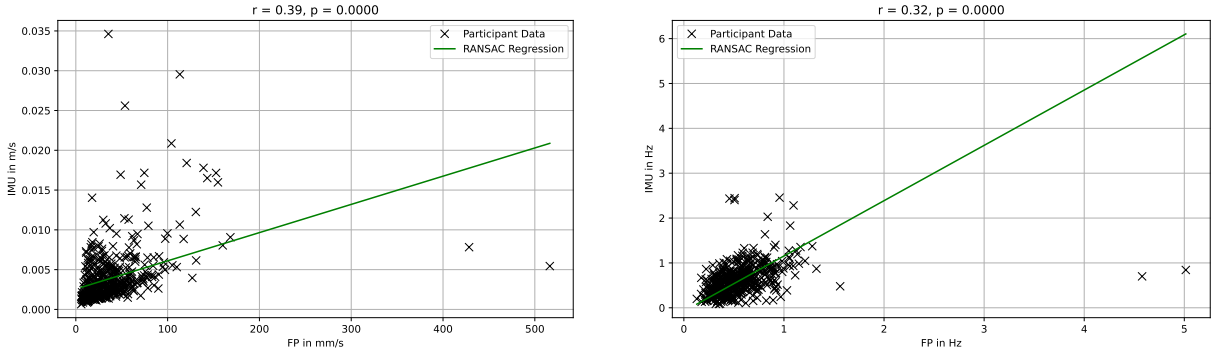
scores of up to 0.47, but high variability across cross-validation folds indicated low generalisability. This outcome can likely be attributed to several factors. First, the binary outcome label for fall risk was based on the FES-I, a self-reported measure of fear of falling, rather than on observed falls or prospectively validated clinical assessments. This may have introduced subjectivity and potential misclassification into the labels, thereby reducing the strength of the learning signal for the models. Second, the underlying data consisted solely of static balance tasks, which may not have provided sufficient discriminatory power to distinguish between groups at risk of falling. This is further supported by the fact that predictive performance remained similarly limited regardless of whether IMU or FP features were used, suggesting that the limitation lies in the data itself rather than the sensing modality. Third, the dataset was unbalanced, with relatively few participants categorized as at risk of falling, which may have further affected the model's generalizability and ability to learn meaningful distinctions. Dealing with missing values that emerged due to participants not completing tasks was also a challenge and required imputation during pre-processing, which may have introduced additional uncertainties. However, the imputation method did not seem to have a noticeable effect on the classification results. A similar pattern was observed in the prediction of cognitive status (MoCA), where the same problems may also have limited the performance of the model. All of these points indicate that the available dataset may not have provided sufficient discriminatory information to reliably distinguish fall risk groups, and that this limitation is more likely due to the study design than to limitations of the IMU or FP modalities themselves. Fall risk prediction using both IMU- and FP–derived sway features did not yield reliable results, indicating that the current data, outcome label and task design were insufficient for robust classification. Nevertheless, the strong agreement between IMU- and FP–derived sway features highlights the potential of ear-worn IMUs as practical tools for balance assessment, even though fall risk prediction remains a challenge with the current data.

These findings underline that while the use of ear-worn IMU-derived sway features for fall risk assessment is feasible, achieving reliable classification outcomes will require more robust, diverse, and longitudinal datasets with validated outcome measures. Future work should aim to collect retrospective fall outcomes through prospective study designs, for example, by recording balance data of participants and using fall diaries or follow-up assessments over several months to establish fall events as ground truth labels for model training. In addition, future studies could benefit from collecting data across multiple time points to capture intra-individual variability and temporal progression of balance impairments. This could be done by exploring modeling strategies that account for individual baseline characteristics, such as personal balance profiles or mobility levels, which can serve as a reference point to detect subtle changes. Combining this data with

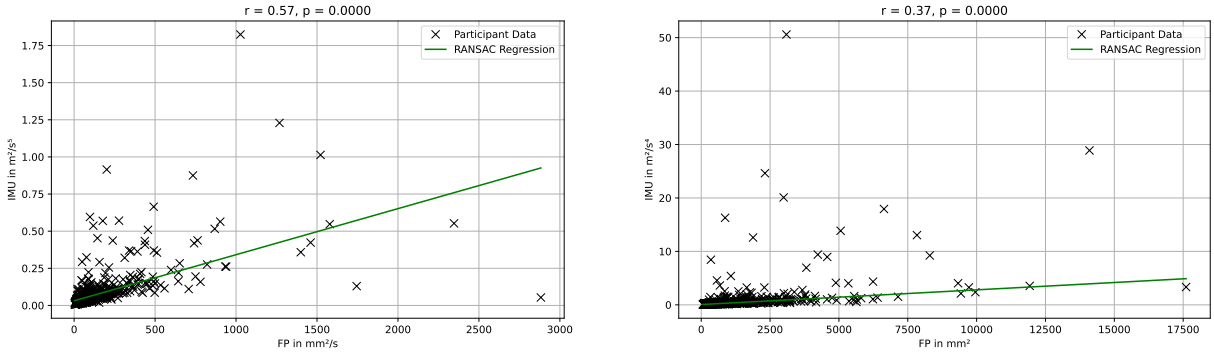
additional health-related measures such as gait patterns, physical activity levels, or neurocognitive screening results may offer a more comprehensive understanding of fall risk. With increased data availability from longitudinal and multi-modal recordings, the resulting data load can be managed by applying more sophisticated modeling approaches beyond traditional machine learning. These might include deep learning architectures capable of capturing complex temporal dynamics in sensor data. Learning directly from raw or minimally pre-processed signals, models such as convolutional networks or recurrent neural networks could improve prediction performance and enable more differentiated assessments. From a clinical and technological perspective, the ability to assess balance and cognitive health using a single, unobtrusive, ear-worn sensor would open new possibilities for remote monitoring and early intervention. Such wearable systems could ultimately support clinicians in identifying individuals with increased fall risk earlier and more efficiently than traditional assessments alone.

Appendix A

Detailed Correlation Results



MV MF

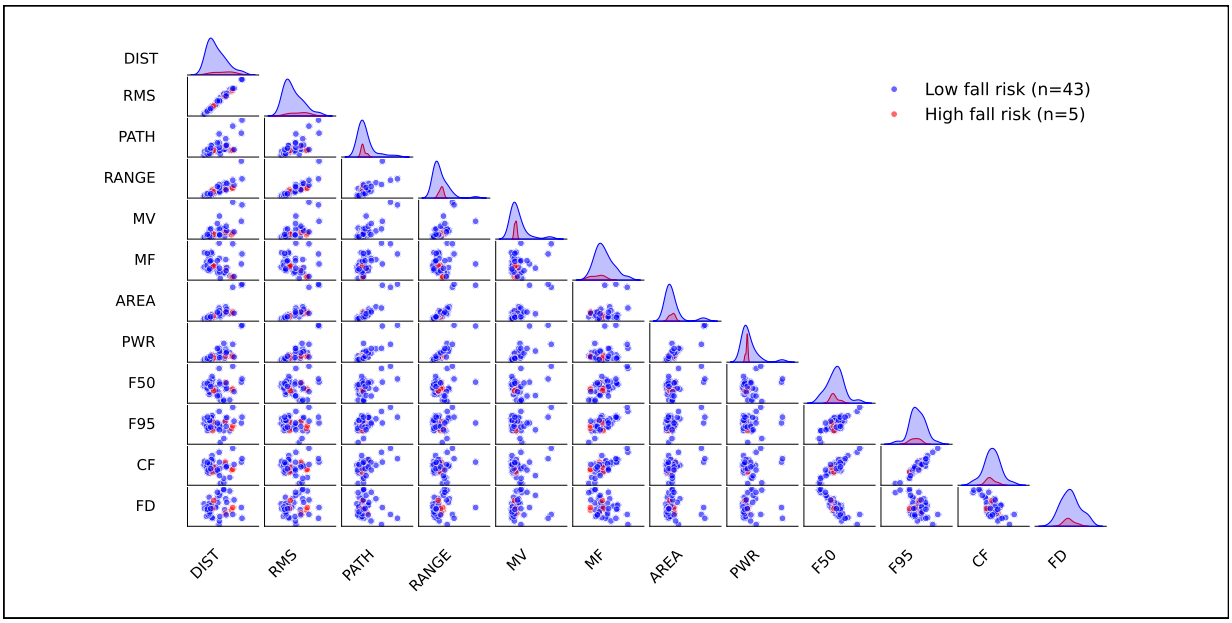


AREA PWR

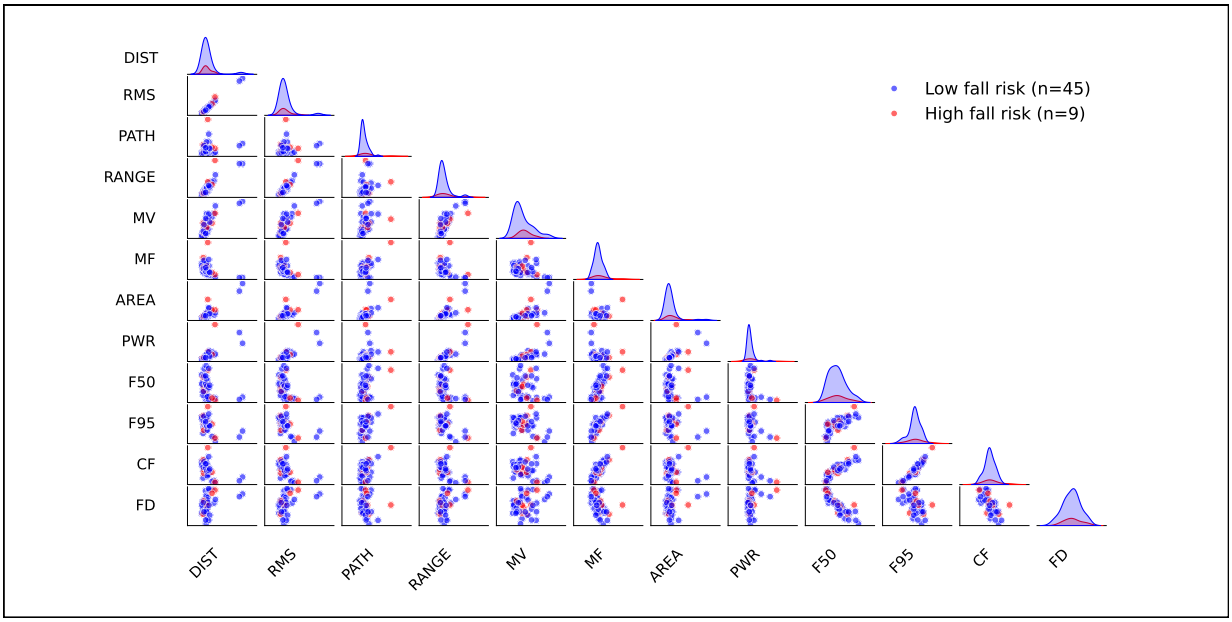
Figure A.1: Full-scale versions of the correlation plots between IMU- and FP–derived features MV, MF, AREA, and PWR across all balance tasks. A regression line fitted using RANSAC is shown for visual reference. These correspond to the zoomed-in plots in Figure 5.5 and 5.6.

Appendix B

Detailed Fall Risk Classification Results

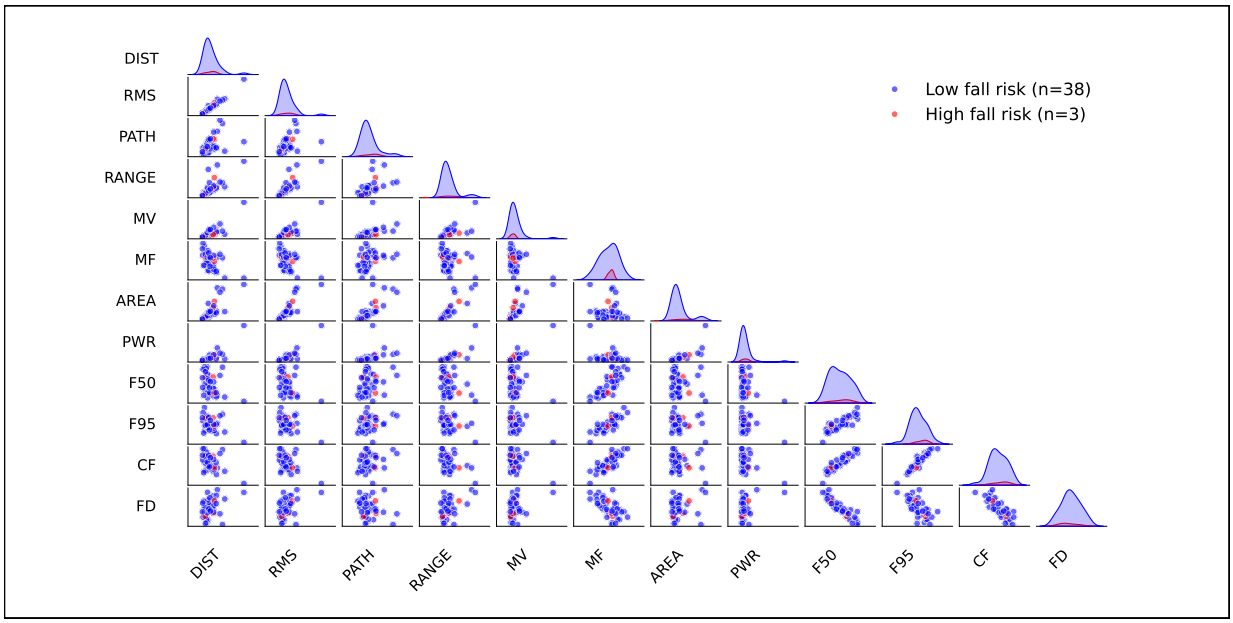


EC Semi

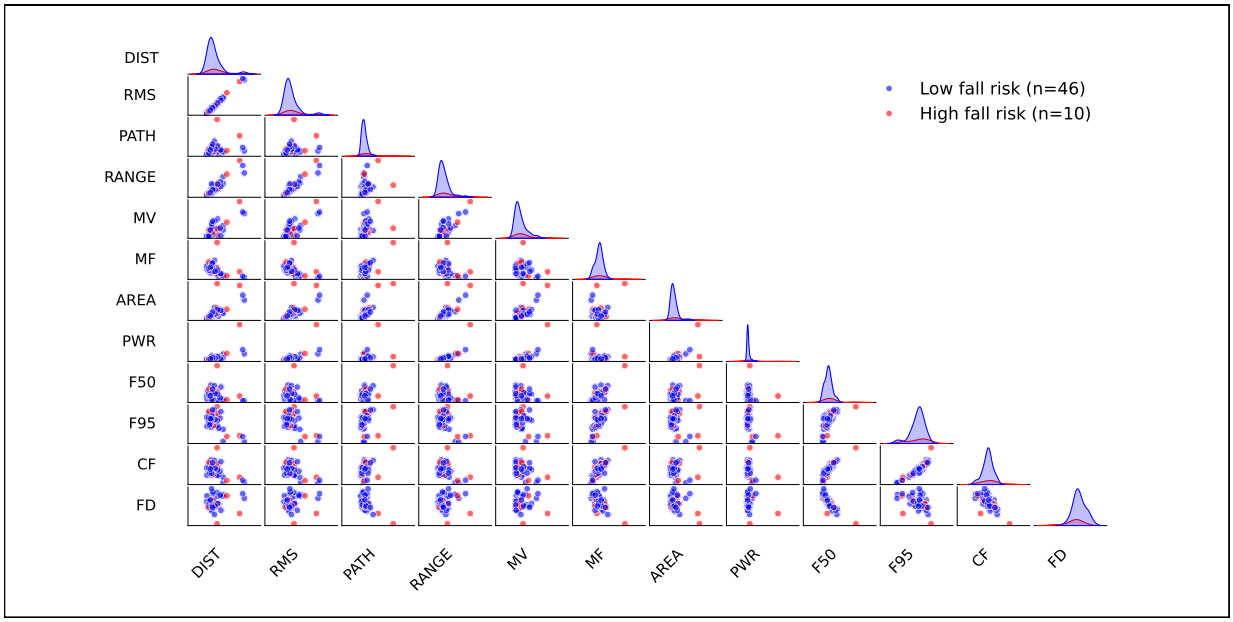


EO semi

Figure B.1: Pairplots of ACC sway metrics of tasksEC semi and EO semi; Low fall risk (FES-I ≤ 22) and high fall risk (FES-I > 22).

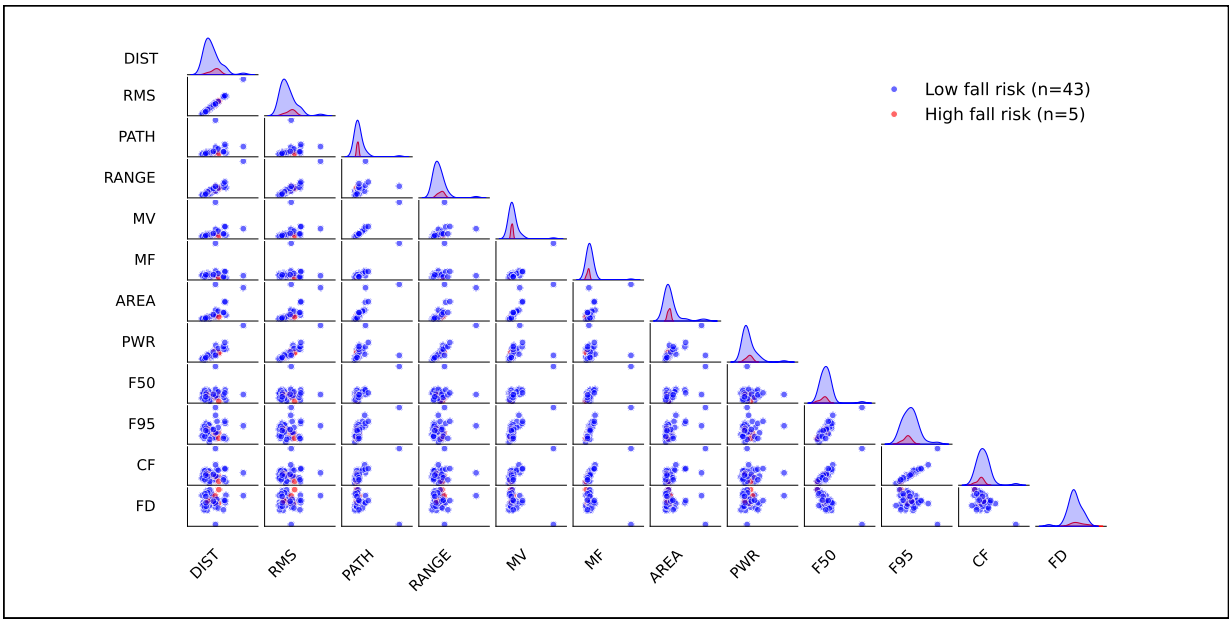


EO tandem

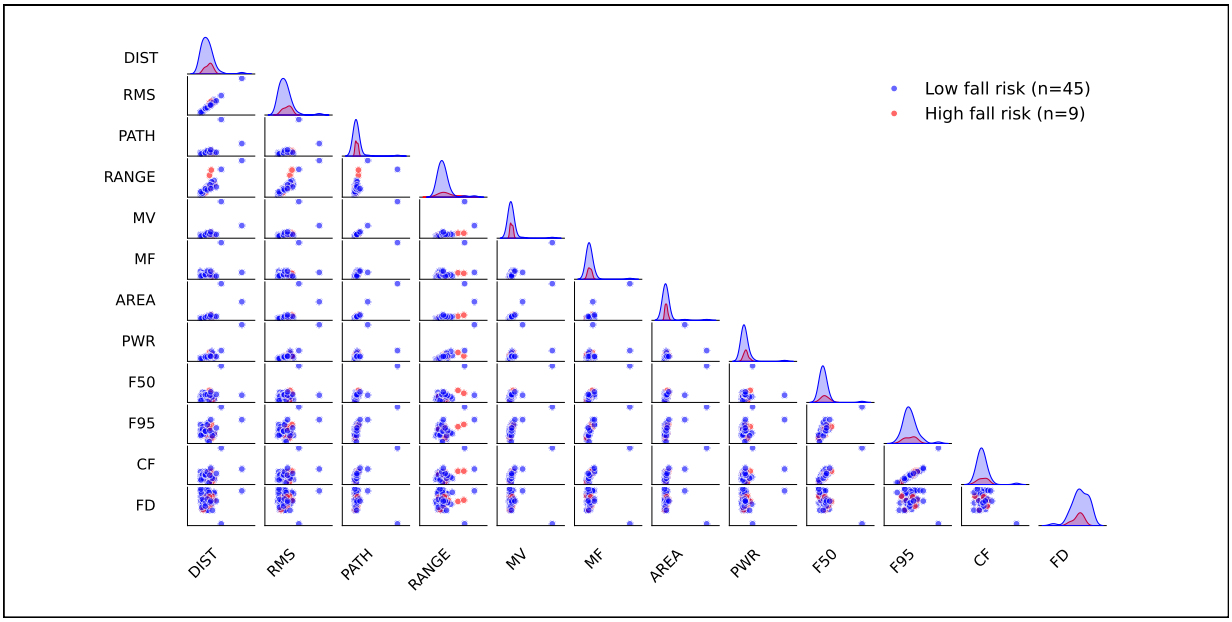


EO pad

Figure B.2: Pairplots of ACC sway metrics of tasks EO tandem and EO pad; Low fall risk (FES-I ≤ 22) and high fall risk (FES-I > 22).

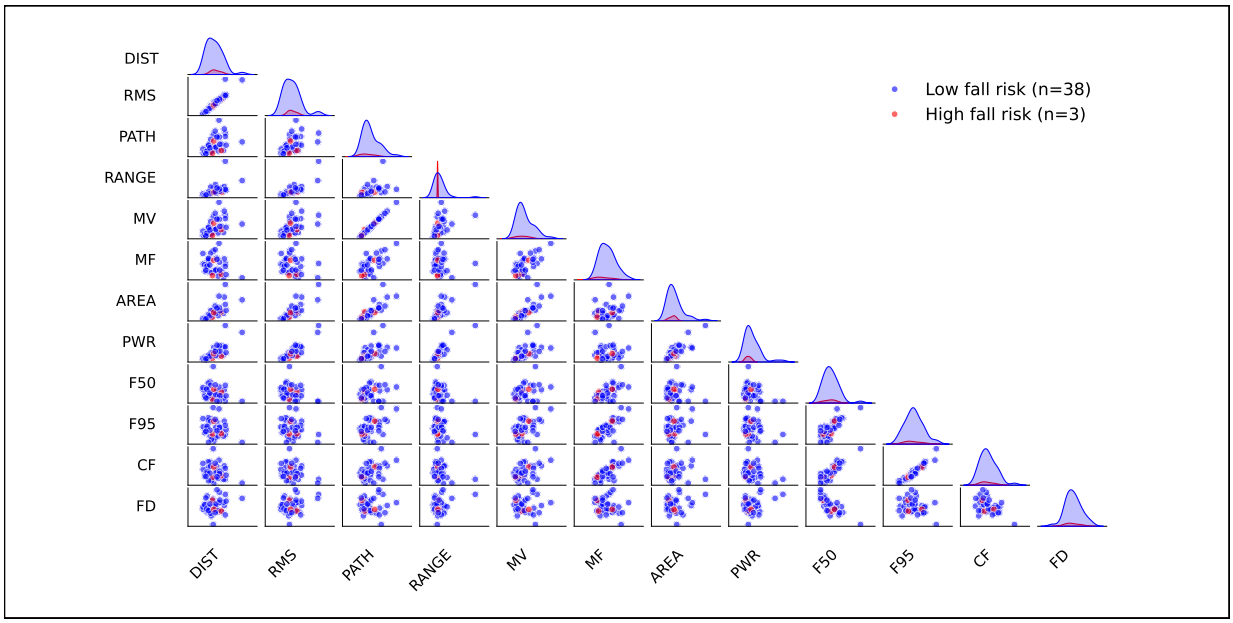


EC Semi

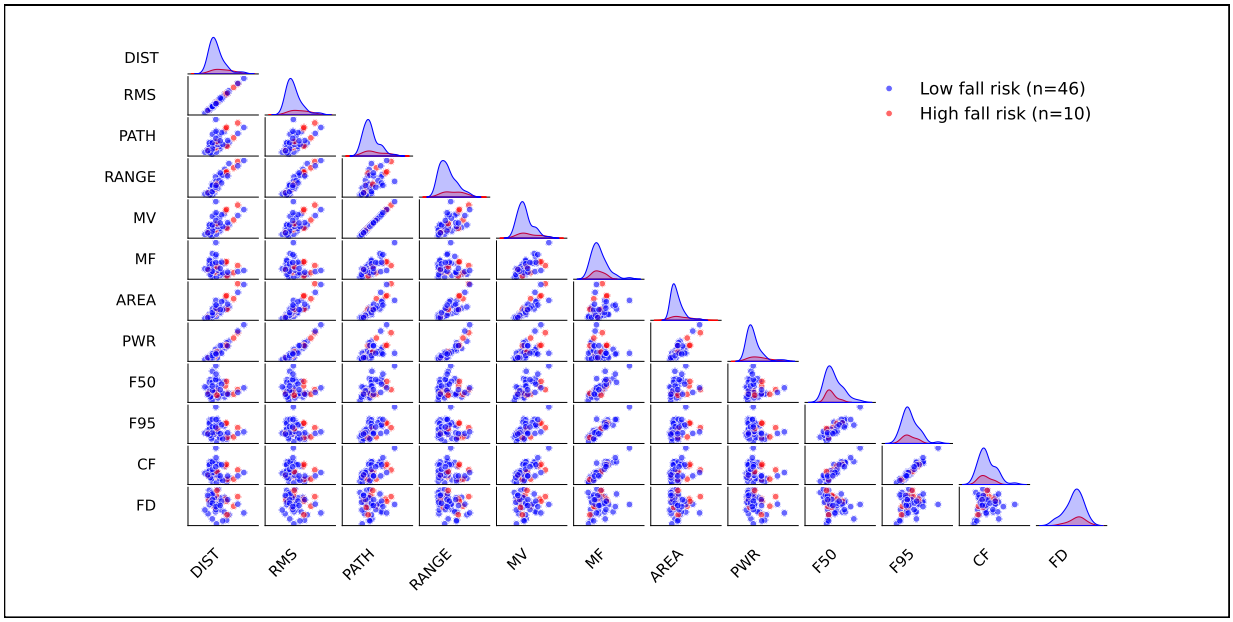


EO semi

Figure B.3: Pairplots of COP sway metrics of tasks EC semi and EO semi; Low fall risk (FES-I ≤ 22) and high fall risk (FES-I > 22).



EO tandem



EO pad

Figure B.4: Pairplots of COP sway metrics of tasks EO tandem and EO pad; Low fall risk (FES-I ≤ 22) and high fall risk (FES-I > 22).

List of Figures

List of Tables

Bibliography

[Agr13] Yuri Agrawal, Bryan K Ward, and Lloyd B Minor. "Vestibular dysfunction: Prevalence, impact and need for targeted treatment". In: *Journal of vestibular research* 23.3 (2013), pp. 113–117.

[Ahm13] Norhafizan Ahmad, Raja Ariffin Raja Ghazilla, Nazirah M Khairi, and Vijayabaskar Kasi. "Reviews on various inertial measurement unit (IMU) sensor applications". In: *International journal of signal processing systems* 1.2 (2013), pp. 256–262.

[All13] Natalie E Allen, Allison K Schwarzel, and Colleen G Canning. "Recurrent falls in parkinson's disease: A systematic review". In: *Parkinson's disease* 2013.1 (2013), p. 906274.

[Apt20] Deborah Apthorp, Alex Smith, Susanne Ilschner, Robin Vlieger, Chandi Das, Christian J Lueck, and Jeffrey CL Looi. "Postural sway correlates with cognition and quality of life in Parkinson's disease". In: *BMJ neurology open* 2.2 (2020).

[Ata11] Louis Atallah, Benny Lo, Rachel King, and Guang-Zhong Yang. "Sensor positioning for activity recognition using wearable accelerometers". In: *IEEE transactions on biomedical circuits and systems* 5.4 (2011), pp. 320–329.

[Azi23] Abdul Aziz, Ravi Karkar, Kan Ding, Jay Harvey, and Phuc Nguyen. "Earables as medical devices: Opportunities and challenges". In: *Adjunct proceedings of the 2023 ACM international joint conference on pervasive and ubiquitous computing & the 2023 ACM international symposium on wearable computing*. (2023), pp. 339–341.

[Ben21] Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. "A comparative analysis of gradient boosting algorithms". In: *Artificial intelligence review* 54 (2021), pp. 1937–1967.

[Ber92] Katherine Berg. "Measuring balance in the elderly: Development and validation of an instrument". PhD thesis. *McGill university*, (1992).

[Bet19] Patricia Bet, Paula C Castro, and Moacir A Ponti. “Fall detection and fall risk assessment in older person using wearable sensors: A systematic review”. In: *International journal of medical informatics* 130 (2019), p. 103946.

[Bin24] VA Binson, Sania Thomas, M Subramoniam, J Arun, S Naveen, and S Madhu. “A review of machine learning algorithms for biomedical applications”. In: *Annals of biomedical engineering* 52.5 (2024), pp. 1159–1183.

[Bis06] Christopher M Bishop and Nasser M Nasrabadi. “Pattern recognition and machine learning”. Vol. 4. 4. *Springer*, (2006).

[Bob24] Lukas Boborzi, Julian Decker, Razieh Rezaei, Roman Schniepp, and Max Wuehr. “Human activity recognition in a free-living environment using an ear-worn motion sensor”. In: *Sensors* 24.9 (2024), p. 2665.

[Bre01] Leo Breiman. “Random forests”. In: *Machine learning* 45 (2001), pp. 5–32.

[Bro13] Harrison James Brown. “Development and validation of an objective balance error scoring system”. PhD thesis. *University of British Columbia*, (2013).

[Cam18] Michelle H Cameron and Ylva Nilsagard. “Balance, gait, and falls in multiple sclerosis”. In: *Handbook of clinical neurology* 159 (2018), pp. 237–250.

[Cha21a] Mayank Arya Chandra and SS Bedi. “Survey on SVM and their application in image classification”. In: *International journal of information technology* 13.5 (2021), pp. 1–11.

[Cha21b] Bahzad Charbuty and Adnan Abdulazeez. “Classification based on decision tree algorithm for machine learning”. In: *Journal of applied science and technology trends* 2.01 (2021), pp. 20–28.

[Che16] Tianqi Chen and Carlos Guestrin. “Xgboost: A scalable tree boosting system”. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. (2016), pp. 785–794.

[Che21] Baoliang Chen, Peng Liu, Feiyun Xiao, Zhengshi Liu, and Yong Wang. “Review of the upright balance assessment based on the force plate”. In: *International journal of environmental research and public health* 18.5 (2021), p. 2696.

[Che22] Manting Chen, Hailiang Wang, Lisha Yu, Eric Hiu Kwong Yeung, Jiajia Luo, Kwok-Leung Tsui, and Yang Zhao. “A systematic review of wearable sensor-based technologies for fall risk assessment in older adults”. In: *Sensors* 22.18 (2022), p. 6752.

[Coh88] Jacob Cohen. “Statistical power analysis for the behavioral sciences”. *Routledge*, (1988).

[Cor95] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine learning* 20 (1995), pp. 273–297.

[Cub24] Diana Maria Cubillos-Arcila, Valéria Feijó Martins, Ana Paula Janner Zanardi, Gustavo Dariva Machado, Daniela Burguêz, Natalia Andrea Gomeñuka, Leonardo Alexandre Peyré-Tartaruga, and Jonas Alex Morales Saute. “Static balance in hereditary spastic paraplegias: A cross-sectional study”. In: *The cerebellum* 23.1 (2024), pp. 162–171.

[Del10] Kim Delbaere, Jacqueline CT Close, A Stefanie Mikolaizak, Perminder S Sachdev, Henry Brodaty, and Stephen R Lord. “The falls efficacy scale-international (FES-I). A comprehensive longitudinal validation study”. In: *Age and ageing* 39.2 (2010), pp. 210–216.

[del16] María del-Río-Valeiras, Pilar Gayoso-Diz, Sofía Santos-Pérez, Marcos Rossi-Izquierdo, Ana Faraldo-García, Isabel Vaamonde-Sánchez-Andrade, Antonio Lirola-Delgado, and Andrés Soto-Varela. “Is there a relationship between short FES-I test scores and objective assessment of balance in the older people with age-induced instability?” In: *Archives of gerontology and geriatrics* 62 (2016), pp. 90–96.

[Del19] Silvia Del Din, Brook Galna, Alan Godfrey, Esther MJ Bekkers, Elisa Pelosin, Freek Nieuwhof, Anat Mirelman, Jeffrey M Hausdorff, and Lynn Rochester. “Analysis of free-living gait in older adults with and without parkinson’s disease and with and without a history of falls: identifying generic and disease-specific characteristics”. In: *The Journals of gerontology: Series A* 74.4 (2019), pp. 500–506.

[Día19] Steven Díaz, Jeannie B Stephenson, and Miguel A Labrador. “Use of wearable sensor technology in gait, balance, and range of motion analysis”. In: *Applied sciences* 10.1 (2019), p. 234.

[Dua10] Marcos Duarte and Sandra MSF Freitas. “Revision of posturography based on force plate for balance evaluation”. In: *Brazilian journal of physical therapy* 14 (2010), pp. 183–192.

[Emp25] EmpkinS. “EmpkinS – Website für den SFB-Antrag Empathokinästhetische Sensorik”. https://empkins.de/. Accessed: 2025-04-22. (2025).

[Exe11] Timothy Exell, David Kerwin, Gareth Irwin, and Marianne Gittoes. "Calculating centre of pressure from multiple force plates for kinetic analyses of sprint running". In: *ISBS-conference proceedings archive*. (2011).

[Fin09] Jonathan T Finnoff, Valerie J Peterson, John H Hollman, and Jay Smith. "Intrarater and interrater reliability of the balance error scoring system (BESS)". In: *Pm&r* 1.1 (2009), pp. 50–54.

[Fri01] Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine". In: *The annals of statistics* 29.5 (2001), pp. 1189–1232.

[Ghi19] Marco Ghislieri, Laura Gastaldi, Stefano Pastorelli, Shigeru Tadano, and Valentina Agostini. "Wearable inertial sensors to assess standing balance: A systematic review". In: *Sensors* 19.19 (2019), p. 4075.

[God08] ACRMDOG Godfrey, Richard Conway, David Meagher, and Gearoid ÓLaighin. "Direct measurement of human movement by accelerometry". In: *Medical engineering & physics* 30.10 (2008), pp. 1364–1386.

[Gra19] Scott T Grafton, Andreas B Ralston, and John D Ralston. "Monitoring of postural sway with a head-mounted wearable device: effects of gender, participant state, and concussion". In: *Medical devices: Evidence and research* (2019), pp. 151–164.

[Hor96] Fay B Horak and Jane M Macpherson. "Postural orientation and equilibrium". In: *Comprehensive physiology* (1996), pp. 255–292.

[How13] Jennifer Howcroft, Jonathan Kofman, and Edward D Lemaire. "Review of fall risk assessment in geriatric populations using inertial sensors". In: *Journal of neuroengineering and rehabilitation* 10 (2013), pp. 1–12.

[How17] Jennifer Howcroft, Edward Lemaire, Jonathan Kofman, and William McILROY. "Elderly fall risk prediction using static posturography". In: *PLOS One* 12 (2017), e0172398.

[Ive13] Grant L Iverson and Michael S Koehle. "Normative data for the balance error scoring system in adults". In: *Rehabilitation research and practice* 2013.1 (2013), p. 846418.

[Jar15] Delaram Jarchi, Benny Lo, Charence Wong, Edmund Ieong, Dinesh Nathwani, and Guang-Zhong Yang. "Gait analysis from a single ear-worn sensor: Reliability and clinical evaluation for orthopaedic patients". In: *IEEE transactions on neural systems and rehabilitation engineering* 24.8 (2015), pp. 882–892.

[Kha21] Prerna Khanna, Tanmay Srivastava, Shijia Pan, Shubham Jain, and Phuc Nguyen. "JawSense: Recognizing unvoiced sound using a low-cost ear-worn system". In: *Proceedings of the 22nd international workshop on mobile computing systems and applications*. (2021), pp. 44–49.

[Kle16] Jochen Klenk, Lars Schwickert, Luca Palmerini, Sabato Mellone, Alan Bourke, Espen AF Ihlen, Ngaire Kerse, Klaus Hauer, Mirjam Pijnappels, Matthis Synofzik, et al. "The FARSEEING real-world fall repository: A large-scale collaborative database to collect and share sensor signals from real-world falls". In: *European review of aging and physical activity* 13 (2016), pp. 1–7.

[Kuh13] Max Kuhn, Kjell Johnson, et al. "Applied predictive modeling". Vol. 26. *Springer*, (2013).

[Lev12] Pazit Levinger, Susannah Wallman, and Keith Hill. "Balance dysfunction and falls in people with lower limb arthritis: factors contributing to risk and effectiveness of exercise interventions". In: *European Review of Aging and Physical Activity* 9 (2012), pp. 17–25.

[Lia23] Yu-Pin Liang and Li-Shan Chou. "Assessment of gait balance control using inertial measurement units—A narrative review". In: *World scientific annual review of biomechanics* 1 (2023), p. 2330006.

[Mad08] Michael E Madden. "Introduction to sectional anatomy". *Lippincott Williams & Wilkins*, (2008).

[Man10] Martina Mancini and Fay B Horak. "The relevance of clinical balance assessment tools to differentiate balance deficits". In: *European journal of physical and rehabilitation medicine* 46.2 (2010), p. 239.

[Man12] Martina Mancini, Arash Salarian, Patricia Carlson-Kuhta, Cris Zampieri, Laurie King, Lorenzo Chiari, and Fay B Horak. "ISway: A sensitive, valid and reliable measure of postural control". In: *Journal of neuroengineering and rehabilitation* 9 (2012), pp. 1–8.

[Mir12] Anat Mirelman, Talia Herman, Marina Brozgol, Moran Dorfman, Elliot Sprecher, Avraham Schweiger, Nir Giladi, and Jeffrey M Hausdorff. "Executive function and falls in older adults: New findings from a five-year prospective study link fall risk to cognition". In: *PLOS One* 7.6 (2012), e40297.

[Mon12] Manuel Montero-Odasso, Joe Verghese, Olivier Beauchet, and Jeffrey M Hausdorff. "Gait and cognition: a complementary approach to understanding brain function and the risk of falling". In: *Journal of the american geriatrics society* 60.11 (2012), pp. 2127–2136.

[Mus24] Isabelle J Museck, Daniel L Brinton, and Jesse C Dean. "The use of wearable sensors and machine learning methods to estimate biomechanical characteristics during standing posture or locomotion: A systematic review". In: *Sensors* 24.22 (2024), p. 7280.

[Nas05] Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. "The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment". In: *Journal of the american geriatrics society* 53.4 (2005), pp. 695–699.

[Ngu16] Anh Nguyen, Raghda Alqurashi, Zohreh Raghebi, Farnoush Banaei-Kashani, Ann C Halbower, and Tam Vu. "A lightweight and inexpensive in-ear sensing system for automatic whole-night sleep stage monitoring". In: *Proceedings of the 14th ACM conference on embedded network sensor systems CD-ROM*. (2016), pp. 230–244.

[Nie83] H. Niemann. "Klassifikation von Mustern". *Springer*, Heidelberg, (1983).

[Nno15] Joseph O Nnodim and Raymond L Yung. "Balance and its clinical assessment in older adults–a review". In: *Journal of geriatric medicine and gerontology* 1.1 (2015), p. 003.

[Noa23a] Alireza Noamani, Negar Riahi, Albert H Vette, and Hossein Rouhani. "Clinical static balance assessment: A narrative review of traditional and IMU-based posturography in older adults and individuals with incomplete spinal cord injury". In: *Sensors* 23.21 (2023), p. 8881.

[Noa23b] Alireza Noamani, Negar Riahi, Albert H Vette, and Hossein Rouhani. "Clinical static balance assessment: a narrative review of traditional and IMU-based posturography in older adults and individuals with incomplete spinal cord injury". In: *Sensors* 23.21 (2023), p. 8881.

[Nob06] William S Noble. "What is a support vector machine?" In: *Nature biotechnology* 24.12 (2006), pp. 1565–1567.

[Pav16] Rita Pavasini, Jack Guralnik, Justin C Brown, Mauro Di Bari, Matteo Cesari, Francesco Landi, Bert Vaes, Delphine Legrand, Joe Verghese, Cuiling Wang, et al. "Short physical performance battery and all-cause mortality: Systematic review and meta-analysis". In: *BMC medicine* 14 (2016), pp. 1–9.

[Pii06] Maarit Piirtola and Pertti Era. "Force platform measurements as predictors of falls among older people–a review". In: *Gerontology* 52.1 (2006), pp. 1–16.

[Pod91] Diane Podsiadlo and Sandra Richardson. "The timed "Up & Go": a test of basic functional mobility for frail elderly persons". In: *Journal of the american geriatrics society* 39.2 (1991), pp. 142–148.

[Pol00] Alexandra S Pollock, Brian R Durward, Philip J Rowe, and John P Paul. "What is balance?" In: *Clinical rehabilitation* 14.4 (2000), pp. 402–406.

[Pol20] Michael Pollind and Rahul Soangra. "Development and validation of wearable inertial sensor system for postural sway analysis". In: *Measurement* 165 (2020), p. 108101.

[Pri96] T.E. Prieto, J.B. Myklebust, R.G. Hoffmann, E.G. Lovett, and B.M. Myklebust. "Measures of postural steadiness: Differences between healthy young and elderly adults". In: *IEEE transactions on biomedical engineering* 43.9 (1996), pp. 956–966.

[Qui21] Flavien Quijoux, Alice Nicolaï, Ikram Chairi, Ioannis Bargiotas, Damien Ricard, Alain Yelnik, Laurent Oudre, François Bertin-Hugault, Pierre-Paul Vidal, Nicolas Vayatis, et al. "A review of center of pressure (COP) variables to quantify standing balance in elderly people: Algorithms and open-access code". In: *Physiological reports* 9.22 (2021).

[Qui86] J. Ross Quinlan. "Induction of decision trees". In: *Machine learning* 1 (1986), pp. 81–106.

[Ras21] Sebastian Raschka. "Model evaluation, model selection, and algorithm selection in machine learning. arXiv 2018". In: *arXiv preprint arXiv:1811.12808* (2021).

[Röd22] Tobias Röddiger, Christopher Clarke, Paula Breitling, Tim Schneegans, Haibin Zhao, Hans Gellersen, and Michael Beigl. "Sensing with earables: A systematic literature review and taxonomy of phenomena". In: *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 6.3 (2022), pp. 1–57.

[Rub06] Laurence Z Rubenstein. "Falls in older people: epidemiology, risk factors and strategies for prevention". In: *Age and ageing* 35.2 (2006), pp. ii37–ii41.

[Saa07] Maytal Saar-Tsechansky and Foster Provost. “Handling missing values when applying classification models”. In: *Journal of machine learning research* 8 (2007), pp. 1625–1657.

[Sai15] Takaya Saito and Marc Rehmsmeier. “The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets”. In: *PLOS One* 10.3 (2015), e0118432.

[Sal17] Santiago J Saldana, Anthony P Marsh, W Jack Rejeski, Jack K Haberl, Peggy Wu, Scott Rosenthal, and Edward H Ip. “Assessing balance through the use of a low-cost head-mounted display in older adults: A pilot study”. In: *Clinical interventions in aging* (2017), pp. 1363–1370.

[Sal18] Joseph P Salisbury, Neha U Keshav, Anthony D Sossong, and Ned T Sahin. “Concussion assessment with smartglasses: Validation study of balance measurement toward a lightweight, multimodal, field-ready platform”. In: *JMIR mhealth and uhealth* 6.1 (2018), e8478.

[Sar21] Iqbal H Sarker. “Machine learning: Algorithms, real-world applications and research directions”. In: *SN computer science* 2.3 (2021), p. 160.

[Sei12] Christina Seimetz, Danica Tan, Riemann Katayama, and Thurmon Lockhart. “A comparison between methods of measuring postrual stability: force plates versus accelerometers”. In: *Biomedical sciences instrumentation* 48 (2012), p. 386.

[Sei23] Ann-Kristin Seifer, Eva Dorschky, Arne Küderle, Hamid Moradi, Ronny Hannemann, and Björn M Eskofier. “EarGait: Estimation of temporal gait parameters from hearing aid integrated inertial sensors”. In: *Sensors* 23.14 (2023), p. 6565.

[Sli19] Salwa O Slim, Ayman Atia, Marwa MA Elfattah, and Mostafa-Sami M Mostafa. “Survey on human activity recognition based on acceleration data”. In: *International journal of advanced computer science and applications* 10.3 (2019).

[Stu08] Daina L Sturnieks, Rebecca St George, and Stephen R Lord. “Balance disorders in the elderly”. In: *Neurophysiologie clinique/clinical neurophysiology* 38.6 (2008), pp. 467–478.

[Sun19] Ruopeng Sun, Katherine L Hsieh, and Jacob J Sosnoff. “Fall risk prediction in multiple sclerosis using postural sway measures: A machine learning approach”. In: *Scientific reports* 9.1 (2019), p. 16154.

[Ull22] Martin Ullrich, Nils Roth, Arne Küderle, Robert Richer, Till Gladow, Heiko Gaßner, Franz Marxreiter, Jochen Klucken, Bjoern M Eskofier, and Felix Kluge. "Fall risk prediction in parkinson's disease using real-world inertial sensor gait data". In: *IEEE journal of biomedical and health informatics* 27.1 (2022), pp. 319–328.

[Vai20] Raju Vaishya and Abhishek Vaish. "Falls in older adults are serious". In: *Indian journal of orthopaedics* 54 (2020), pp. 69–74.

[Van12] Stef Van Buuren and Stef Van Buuren. "Flexible imputation of missing data". Vol. 10. *CRC press Boca Raton, FL*, (2012).

[Vij21] Vini Vijayan, James P Connolly, Joan Condell, Nigel McKelvey, and Philip Gardiner. "Review of wearable devices and data collection considerations for connected health". In: *Sensors* 21.16 (2021), p. 5589.

[Vir20] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. "SciPy 1.0: Fundamental algorithms for scientific computing in Python". In: *Nature methods* 17.3 (2020), pp. 261–272.

[Wan24] Jixian Wang, Yongfang Li, Guo-Yuan Yang, and Kunlin Jin. "Age-related dysfunction in balance: a comprehensive review of causes, consequences, and interventions". In: *Aging Dis* (2024), pp. 2024–0124.

[Wei13] Aner Weiss, Marina Brozgol, Moran Dorfman, Talia Herman, Shirley Shema, Nir Giladi, and Jeffrey M Hausdorff. "Does the evaluation of gait quality during daily life provide insight into fall risk? A novel approach using 3-day accelerometer recordings". In: *Neurorehabilitation and neural repair* 27.8 (2013), pp. 742–752.

[Wer23] Lars Werntz and Eva Greiner. "Physiologie und Anatomie; Grundlagen der Humanbiologie für Pharmaziestudierende". *Govi-Verlag*, (2023).

[Whi05] Tim D. White and Pieter A. Folkens. "Chapter 6 - Anatqmical terminqlqgy". In: "The human bone manual". *Academic press*, San Diego, (2005), pp. 67–74.

[Wil23] Drew Wilimitis and Colin G Walsh. "Practical considerations and applied examples of cross-validation for model development and evaluation in health care: tutorial". In: *Jmir ai* 2 (2023), e49023.

[Woo09] John C Woolcott, Kathryn J Richardson, Matthew O Wiens, Bhavini Patel, Judith Marin, Karim M Khan, and Carlo A Marra. “Meta-analysis of the impact of 9 medication classes on falls in elderly persons”. In: *Archives of internal medicine* 169.21 (2009), pp. 1952–1960.

[Yar05] Lucy Yardley, Nina Beyer, Klaus Hauer, Gertrudis Kempen, Chantal Piot-Ziegler, and Chris Todd. “Development and initial validation of the falls efficacy scale-international (FES-I)”. In: *Age and ageing* 34.6 (2005), pp. 614–619.

[Zan24] Kirsten Zantvoort, Barbara Nacke, Dennis Görlich, Silvan Hornstein, Corinna Jacobi, and Burkhardt Funk. “Estimation of minimal data sets sizes for machine learning predictions in digital mental health interventions”. In: *npj digital medicine* 7.1 (2024), p. 361.

[Zha22] Shibo Zhang, Yaxuan Li, Shen Zhang, Farzad Shahabi, Stephen Xia, Yu Deng, and Nabil Alshurafa. “Deep learning in human activity recognition with wearable sensors: A review on advances”. In: *Sensors* 22.4 (2022), p. 1476.

Appendix C

Acronyms

ABC Activities-specific Balance Confidence

ACC Accelerometer

AP Antero-Posterior

AREA Sway Area per Second

BBS Berg Balance Scale

BESS Balance Error Scoring System

CF Centroidal Frequency

COP Center of Pressure

DIST Mean Distance from the Center of the Trajectory

DHI Dizziness Handicap Inventory

EC Eyes Closed

EO Eyes Open

FD Frequency Dispersion

FES-I Falls Efficacy Scale-International

FN False Negative

FP Force Plate

FP False Positive

FPR False Positive Rate

F50 Median Frequency

F95 95% Power Frequency

GRF Ground Reaction Force

IMU Inertial Measurement Unit

KNN k-Nearest Neighbors

MF Mean Frequency

ML Medio-Lateral

MoCA Montreal Cognitive Assessment

MV Mean Velocity

oBESS Objective Balance Error Scoring System

PATH Total Sway Path Length

PWR Total Power

PSD Power Spectral Density

RANGE Range of Displacement

RANSAC Random Sample Consensus

RF Random Forest

RMS Root Mean Square

ROC Receiver Operating Characteristic

ROC-AUC Receiver Operating Characteristic - Area under the Curve

Short FES-I Short Falls Efficacy Scale–International

SPPB Short Physical Performance Battery

SVM Support Vector Machine

SMOTE Synthetic Minority Over-sampling Technique

SI Superior-Inferior

TP True Positive

TN True Negative

TPR True Positive Rate

TUG Timed Up and Go

VRHMD Virtual Reality Head-Mounted Display

XGBoost Extreme Gradient Boosting